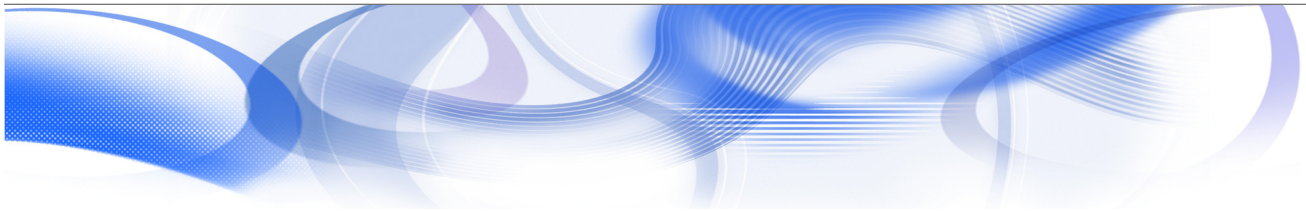
JRC Scientific and Technical Reports



In Silico Prediction of Physicochemical Properties

John Dearden[a] and Andrew Worth[b]

[a] QSTAR Consulting, Helsby, Cheshire, England

[b] European Chemicals Bureau, Institute for Health & Consumer Protection, European Commission - Joint Research Centre, Ispra, Italy

EUR 23051 EN - 2007

JRC EUROPEAN COMMISSION

ihcp Institute for Health and Consumer Protection

The mission of the IHCP is to provide scientific support to the development and implementation of EU policies related to health and consumer protection.

The IHCP carries out research to improve the understanding of potential health risks posed by chemical, physical and biological agents from various sources to which consumers are exposed.

European Commission
Joint Research Centre
Institute for Health and Consumer Protection

Contact information
Address: European Chemicals Bureau TP 581
E-mail: andrew.worth@ec.europa.eu
Tel.: +39 0332 789566
Fax: +39 0332 786717

http:// http://ecb.jrc.it/QSAR
http://www.jrc.cec.eu.int

JRC 42061

EUR 23051 EN

ISSN 1018-5593

Luxembourg: Office for Official Publications of the European Communities

Printed in Italy

Quality control insert

	Name	*Signature*	*Date*
Report Prepared by:	Andrew Worth		03/12/07
Reviewed by: **(Scientific level)**	Tatiana Netzeva		10/12/07
Approved by: **(Head of Unit)**	Steven Eisenreich		
Final approval: **(IHCP Director)**	Elke Anklam		

ABSTRACT

This report provides a critical review of computational models, and in particular (quantitative) structure-property relationship (QSPR) models, that are available for the prediction of physicochemical properties. The emphasis of the review is on the usefulness of the models for the regulatory assessment of chemicals, particularly for the purposes of the new European legislation for the Registration, Evaluation, Authorisation and Restriction of CHemicals (REACH), which entered into force in the European Union (EU) on 1 June 2007.

It is estimated that some 30,000 chemicals will need to be further assessed under REACH. Clearly, the cost of determining the toxicological and ecotoxicological effects, the distribution and fate of 30,000 chemicals would be enormous. However, the legislation makes it clear that testing need not be carried out if adequate data can be obtained through information exchange between manufacturers, from in vitro testing, and from in silico predictions.

The effects of a chemical on a living organism or on its distribution in the environment is controlled by the physicochemical properties of the chemical. Important physicochemical properties in this respect are, for example, partition coefficient, aqueous solubility, vapour pressure and dissociation constant. Whilst all of these properties can be measured, it is much quicker and cheaper, and in many cases just as accurate, to calculate them by using dedicated software packages or by using (QSPRs). These *in silico* approaches are critically reviewed in this report.

LIST OF ABBREVIATIONS

AIT	Auto-ignition temperature
ANN	Artifical Neural Network
Dow	Octanol-water distribution coefficient
EC	European Commission
ECB	European Chemicals Bureau
ECETOC	European Centre for Ecotoxicology and Toxicology of Chemicals
EPA	Environmental Protection Agency
F	Fugacity ratio or Fischer Statistic
GSE	General Solubility Equation
Ka	Acid Dissociation Constant (often as pKa)
Koc	Organic carbon coefficient – measure of soil sorption
Kom	Organic matter coefficient – measure of soil sorption
Kow	Octanol-water partition coefficient
MAE	Mean absolute error
MP	Melting point
MW	Molecular Weight
OECD	Organisation for Economic Cooperation and Development
Q^2	Cross-validated R^2 - measure of predictive ability
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure-Property Relationship
R^2	Coefficient of determination – measure of goodness-of-fit
REACH	Registration Evaluation and Authorisation of Chemicals
RMS	Root mean square
s	Standard error of the estimate
Saq	Aqueous solubility
SMILES	Simplified Molecular Line Entry system
T_b	Boiling point
T_F	Flash point
T_m	Melting point
VP	Vapour pressure

CONTENTS

1. Introduction

REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) is the new chemicals legislation for the European Union, which aims to ensure that existing chemicals made and used in the EU have adequate toxicity data. Currently there is a grave paucity of toxicological information concerning these chemicals. Even for chemicals produced in high volume (> 1000 tonne/year) there are remarkably few toxicity data; 3% have adequate toxicity data, 11% have a minimal base set of data, 65% have very few data and 21% have no toxicity data at all. Some 30,000 chemicals are covered by REACH, which was adopted by the European Parliament on 13 December 2006, and which entered into force on 1 June 2007 [EC 2006a, 2006b].

Clearly, the cost of toxicity testing of 30,000 chemicals would be enormous, when one considers that to test one chemical for carcinogenicity takes two years and costs over 5 million Euro, and that other toxicity tests, for example for mutagenicity, teratogenicity, allergenicity and endocrine disruption may also need to be carried out. However, the legislation makes it clear that animal testing should not be carried out unless other toxicological data are inadequate. Such data may be obtained through information exchange between manufacturers, from in vitro testing, and from *in silico* predictions.

The effects of a chemical on living organisms and its distribution in the environment are controlled by the physicochemical properties of the chemical. Important physicochemical properties in this respect are, for example, partition coefficient, aqueous solubility, vapour pressure and dissociation constant. Whilst all of these properties can be measured, it is much quicker and cheaper, and in many cases just as accurate, to calculate them using dedicated computer software or by the use of quantitative structure-property relationships (QSPRs).

Under the REACH legislation, the physicochemical properties required are:

Under Annex V, for substances at a supply level of ≥ 1 tonne/year: octanol-water partition coefficient, aqueous solubility, melting/freezing point, boiling point, vapour pressure, relative density, surface tension, flash point, auto-ignition (self-

ignition) temperature, flammability (including pyrophoric properties and flammability on contact with water), explosive properties, oxidising properties and granulometry (particle size distribution).

Under Annex VI, for substances at a supply level of ≥ 10 tonne/year: adsorption/desorption screening (soil sorption).

Under Annex VII, for substances at a supply level of ≥ 100 tonne/year: dissociation constant, viscosity, stability in organic solvents.

Of the above properties, flammability, explosive properties, oxidising properties, granulometry and stability in organic solvents are not amenable to calculation, and hence are not discussed in this report. The report does, however, include a section on the calculation of the air-water partition coefficient (Henry's law constant), since this is an important property that controls the environmental distribution of a chemical.

2. Introduction to quantitative structure-property relationships (QSPRs)

Every property (physical, chemical and biological) of a chemical compound depends on the molecular structure of that compound. A simple example of this is the fact that melting point increases and water solubility (S_{aq}) decreases as one ascends an homologous series. In this example, the change in structure along the homologous series could be represented by an increase in carbon number or molecular weight (MW). Thus one would expect a correlation between water solubility and molecular weight, within the homologous series:

$$\log S_{aq} = a\,MW + c \qquad (1)$$

where a and c are constants. The logarithm of the property value is often used, firstly because property values can span many orders of magnitude, and secondly because such properties are usually rate constants or equilibrium constants, and as such their logarithm is proportional to the free energy change of the reaction or equilibrium. The

correlations are known as linear free energy relationships, or more commonly as quantitative structure-property relationships (QSPRs). The term quantitative structure-activity relationship (QSAR) is sometimes used instead of QSPR, but strictly this should be reserved for the correlation of biological activities.

Introductions to QSPRs have been published by Walker et al [2003] and by Dearden and Cronin [2006]. The application of QSPRs to environmental assessment has been discussed by Borman [1990], Russom et al [2003] and Walker et al [2003]. Two very useful documents for those wishing to use QSPRs in environmental assessment are OECD Environment Monograph No. 67 [1993] and ECETOC Technical Report No. 74 [1998].

Generic guidance on the regulatory use of QSARs and QSPRs under REACH is given in ECB [2007].

2.1 The value of QSPRs

QSPRs are valuable in two main ways. Firstly, they can be used to predict the property in question of compounds that were not used to develop the QSPR (i.e. that were not in the training set). In order for this to be valid, the compounds should be similar to those used in the training set. That is, they should be within what is called the "applicability domain" of the QSPR [Netzeva et al 2005]. Thus, if equation (1) was a QSPR developed with a training set of substituted benzenes, it would not give accurate predictions for aliphatic alcohols. Again, if the MW range of the benzene derivatives used in the training set was 100 to 250, the QSPR would probably not give an accurate prediction for a benzene derivative with MW = 500. Netzeva et al [2005], Dimitrov et al [2005], Nikolova-Jeliazkova and Jaworska [2005] and Schultz et al [2007] have discussed the determination of applicability domains. A recent useful critical assessment of QSARs in ecotoxicological risk assessment is given by de Roode et al [2006].

The second way that QSPRs can be useful is that the descriptor(s) that best model the property in question may throw some light on the mechanism that governs the

property. For example, sorption of chemicals on soils is found to be correlated well with the octanol-water partition coefficient. One could therefore deduce that the sorption mechanism is a partitioning process between water and the hydrophobic surface of soil particles. However, one must be careful not to place too much reliance on such interpretations, since a correlation is not proof of cause-and-effect.

The structure descriptors (that is, the terms on the right hand side of the correlation) are usually structural or physicochemical properties. There are thousands of such descriptors available, and numerous software programs are available for their calculation. Descriptor values are also available in many books and compendia [e.g. Hansch & Leo 1995].

2.2 Development of a QSPR

In order to develop a QSPR, one firstly needs property values, such as aqueous solubilities, for a series of compounds. The series of compounds may be a congeneric series, such as a series of substituted phenols, or it may be a very diverse set of chemicals. Better correlations are generally obtained for congeneric series, because even for a simple physicochemical process like dissolution, different mechanisms of dissolution can be envisaged for different classes of compounds. It is nevertheless true that, especially for physicochemical properties, good correlations are often obtained for very diverse training sets of compounds. One such QSPR, for aqueous solubility of a large and diverse set of organic chemicals, was developed by Abraham and Le [1999] using their so-called solvatochromic descriptors:

$$\log S_{aq} = 0.518 - 1.004\,R + 0.771\,\pi^H + 2.168\,\Sigma\alpha^H + 4.238\,\Sigma\beta^H - 3.362\,\Sigma\alpha^H.\Sigma\beta^H - 3.987\,V_X \quad (2)$$

$$n = 659 \quad R^2 = 0.920 \quad s = 0.557$$

where R = excess molar refractivity (a measure of polarisability), π^H = a polarity/polarisability term, $\Sigma\alpha^H$ and $\Sigma\beta^H$ = sums of hydrogen bond donor and acceptor abilities respectively, and V_X = McGowan characteristic molecular volume. All of these terms can be calculated with the Absolv-2 software (see Table 1). All the

descriptor values cover approximately the same ranges, so the magnitude of the coefficient of a descriptor indicates its contribution to aqueous solubility. Hence one can see from equation (2) that the two most important descriptors are hydrogen bond acceptor ability and molecular size.

2.3 QSPR statistics

The statistics of the correlation are given after the equation. The number (n) of compounds in the training set was 659; the multivariate coefficient of determination (R^2) is a measure of the fraction of the variation in log (solubility) that is described by the QSPR (note that when a single descriptor is used to model a property, the coefficient of determination is written as r^2); in this case, the QSPR describes 92% of the variation of solubility; the standard error of the estimate (s) gives an indication of the accuracy with which the aqueous solubility of the training set compounds is modelled. In this case, the standard error of the estimate is close to the experimental error in the measurement of aqueous solubility [Katritzky et al 1998], so the QSPR models the data well. Additional statistics, not reported in this case, are the standard errors on each coefficient, which can give an indication of whether or not a particular descriptor contributes significantly to the correlation; if the standard error of a coefficient is close to the value of the coefficient itself, the descriptor contributes little, and should be discarded.

To safeguard against chance correlations, it is recommended [Topliss & Costello 1972] that the ratio of training set compounds to descriptors in the QSPR should be at least 5:1.

2.4 Descriptor selection

Descriptors are generally selected in one of two ways. Firstly, one can "guess" what descriptors might best model the property in question, and use them to derive the QSPR. However, if one's guess is wrong, then a good QSPR will not be obtained. A more commonly used approach is to generate a large number of descriptors (perhaps several hundred) and use a statistical method such as step-wise regression or genetic

algorithm to select the "best" descriptors, i.e. those that give the best correlation with the property in question. This procedure is, however, subject to a higher risk of chance correlations occurring [Topliss & Edwards 1979]

2.5 Predictive ability of a QSPR

Even if a good QSPR is obtained, it may not be a good predictor of the property in question for compounds not in the training set. Hence some measure of predictive ability is required. The best way for predictivity to be assessed is to use the QSPR to predict the property in question for a number of compounds that were not used in the training set, but for which the measured value of the property is known; such a set of compounds is called a test set. The test set compounds must be reasonably similar to those of the training set; that is, they must lie within the applicability domain of the QSPR. This is often achieved by dividing the total number of compounds into two groups; the larger group forms the training set, and the smaller group (typically 5 – 50% of the total) forms the test set. If the standard error for the test set is much larger than that for the training set, then the QSPR does not have good predictivity, and it should not be used for predictive purposes.

If the total number of compounds is small, then it may not be practicable to split it into training and test sets. In that case a procedure called internal cross-validation can be used, whereby each compound in turn is deleted from the training set, the QSPR is developed with the remaining compounds, and is used to predict the property value of the omitted compound. That compound is then returned to the training set and a second compound is deleted, and so on until every compound has been left out in turn. A cross-validated R^2 value, called Q^2, is then calculated, which is an indicator of the internal predictivity of the QSPR. It is, however, not considered to be as good an indicator as is obtained using an external test set. Walker et al [2003] have proposed that an indicator of good predictivity is that Q^2 should not be more than 0.3 lower than R^2, whilst Eriksson et al [2003] have proposed a minimal acceptable value of 0.5 for Q^2.

2.6 Prediction software

Numerous software programs are available for the prediction of physicochemical and other properties of environmental and/or health interest. Some are freely accessible online, some are freely downloadable from a website, whilst others have to be purchased. The availability of such software is given in Table 1. ECETOC [2003] have examined the performance of five of these programs, namely Episuite, ASTER, SPARC, ACD/Labs and PREDICT. For some software, the input format is SMILES (simplified molecular input line entry system). SMILES is extremely easy to learn, and a tutorial can be found at www.daylight.com/smiles/smiles-intro.html. It is recommended that, given availability of appropriate software, property predictions be made in this way, as they can be obtained very quickly. It is recommended that at least three predictions be obtained if possible. For example, the experimental log K_{ow} value for chloramphenicol is 1.14. The on-line AlogPS software (see Table 1) calculates a value of 1.15, whilst the on-lineChemSilico software (see Table 1) and the freely downloadable KOWWIN software (part of the Episuite software; see Table 1) both calculate a value of 0.92. A predicted log K_{ow} value within 0.3 – 0.4 log unit of a good measured value is considered acceptable. The relevant properties predicted by each software program are shown in Table 2.

The question arises as to whether, if a measured property value is available, it should always be used in preference to a calculated value. It should be borne in mind that experimental values are also subject to error. For example, it is generally accepted that the mean experimental error on log K_{ow} values is about 0.3 log unit. This is about the same as the mean error of the best predicted log K_{ow} values. Since property prediction software is generally developed on training sets of several thousand very diverse chemicals, it can be assumed that the applicability domain of such software is very extensive. (Applicability domains are currently not usually available for commercial software.) The added precaution of using predictions from at least three software programs should ensure that a mean predicted property is just as acceptable as a measured value. Of course, if a measured value is available, it should not be ignored.

However, if there is no appropriate software, or if it is too expensive, an appropriate QSPR should be selected. Tetko et al [2006] have discussed the accuracy of

prediction of properties such as log K_{ow}. All the software can be run in batch mode, except where indicated.

Table 1. Software for the prediction of physicochemical properties

Software	Availability	Website address
Absolv-2	Purchase	www.ap-algorithms.com
ACD/Labs[a]	Purchase	www.acdlabs.com
Admensa	Purchase	www.inpharmatica.com
ADME Boxes	Purchase	www.ap-algorithms.com
ADMET Predictor[a]	Purchase	www.simulationsplus.com
ASTER[b]	Not available for public use	www.epa.gov
ChemAxon	Purchase	www.chemaxon.com
ChemOffice	Purchase	www.cambridgesoft.com
ChemProp	Not known	www.ufz.de/index.php?en=6738
ChemSilico	Purchase[c]	www.chemsilico.com
ClogP	Purchase	www.daylight.com
Episuite	Freely downloadable	www.epa.gov/oppt/exposure/pubs/episuitedl.htm
Molecular Modeling Pro	Purchase	www.chemsw.com
Pallas	Purchase	www.compudrug.com
Pipeline Pilot	Purchase	www.scitegic.com
PREDICT	Purchase	mwsoftware.com/dragon/
ProPred	Consortium members only	www.capec.kt.dtu.dk
QikProp	Purchase	www.schrodinger.com
SPARC[d]	Free on-line	ibmlc2.chem.uga.edu/sparc
TSAR	Purchase	www.accelrys.com
VCCLAB	Free on-line	www.vcclab.org

[a]Aqueous solubility module predicts intrinsic solubility, solubility in pure water and solubility at user-specified pH

[b]Aster is currently not available for public use. It is hoped that it will at some point be available on WWW. It is understood that OECD may also add it to their QSAR Toolbox.

[c]Log K_{ow} and aqueous solubility predictions available free on-line at www.logp.com, but not in batch mode

[d]Not available in batch mode

Table 2. Physicochemical properties estimated by commercially and freely available software

Software	Log K_{ow}	Aqueous solubility	pKa	Melting point	Boiling point	Vapour pressure	Henry's law constant	Relative density	Viscosity	Surface tension	Flash point	Soil sorption
Absolv-2	✓	✓				✓	✓					✓
ACD/Labs	✓	✓	✓		✓	✓		✓		✓	✓	
ADME Boxes	✓	✓	✓				✓					✓
ADMET Predictor	✓	✓	✓									
Admensa	✓	✓										
ASTER	✓	✓	✓		✓	✓	✓					
ChemAxon	✓		✓									
ChemOffice	✓			✓	✓							
ChemProp	✓	✓		✓	✓	✓	✓		✓			
ChemSilico	✓	✓	✓									
ClogP	✓											
Episuite	✓	✓		✓	✓	✓	✓					✓
Molecular Modeling Pro	✓	✓		✓	✓	✓		✓	✓			
Pallas	✓		✓									
Pipeline Pilot	✓	✓	✓									
Predict				✓	✓	✓		✓	✓	✓		
ProPred	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	
QikProp	✓	✓	✓									
SPARC	✓	✓	✓		✓	✓	✓	✓				
TSAR	✓											
VCCLAB	✓	✓	✓									

2.7 Steps in the selection of a QSPR for predictive purposes

1. Find a QSPR that is appropriate to the chemical(s) whose property you wish to predict; this is done by searching the scientific literature to find relevant papers or books, or by searching the internet. If the QSPR training set is available, check to see whether similar chemicals to yours are present. If the training set is not available, find out what the authors say about the nature and range of chemicals used in the training set. If that information is not available, find out whether the authors give the ranges of the descriptors used in the

training set (e.g. a log K_{ow} range of 1 – 6). If none of the above information is available, it is recommended that the QSPR is not used.

2. If the QSPR is appropriate, check that the statistics are acceptable. The R^2 value should preferably be > 0.9, and the Q^2 value > 0.6. The standard error (or other measure of error such as RMS error) should be close to the error on experimental measurements. If an external test set has been used, the standard error of the test set results should be similar to that of the training set. For example, the experimental error on aqueous solubility measurements is about 0.5 – 0.6 log unit, so a prediction error of, say, > 0.8 would be unacceptable. Conversely, do not use a QSPR with a standard error much lower than experimental error; this indicates that the QSPR has over-fitted the data, and its predictivity will be poor. If the standard error of each descriptor coefficient is given, check that the standard error is considerably lower than the value of the coefficient, otherwise the descriptor will be of low significance (p value > 0.05, where p is the probability that the descriptor has been selected by chance).

3. Given acceptable statistics, check that the descriptors used in the QSPR are readily available from books, papers, the internet or from accessible and affordable software. If so, proceed with the property prediction.

2.8 Consensus modelling

Some QSPR predictions and software programs are better than others. However, even the good ones do not yield perfect predictions. It is therefore always best, provided that it is practicable, to obtain property predictions from at least three different methods. In that way one can see whether one prediction is very different from the others, and should perhaps be discarded. This is exemplified by four separate software predictions of the aqueous solubility (log S_{aq}, with S_{aq} in mol L^{-1}) of three different chemicals. All the software programs have been tested [Dearden 2006] and found to give good predictions overall. It should be borne in mind that the average experimental error on aqueous solubility measurements is about ± 0.6 log unit.

Table 3. Software predictions of aqueous solubility

	Atropine	Caffeine	Butylparaben
Measured	- 2.18	- 1.02	- 2.96
Software no. 1	- 1.87	- 1.87	- 3.09
Software no. 2	- 2.06	- 0.65	- 3.05
Software no. 3	- 1.01	- 0.27	- 3.07
Software no. 4	- 2.03	- 0.56	- 2.58

For atropine, it is clear that three programs give similar predictions, well within the experimental error, whereas software no. 3 gives a poor prediction. The mean of the three good predictions is – 1.99, which is only 0.2 log unit different from the experimental value of – 2.18. However, it may be noted that even if the poor prediction from software no. 3 is included, the mean predicted value is – 1.74, which is still within the experimental error of ± 0.6 log unit from the measured value.

For caffeine, there is a considerable divergence of predicted values, indicating that the solubility of this compound is difficult to predict. Only two of the four predictions are within the experimental error of ± 0.6 log unit, but the mean of all four predictions is – 0.84, which is well within the experimental error. This example really emphasises the value of consensus modelling.

Butylparaben has a simpler chemical structure than those of atropine and caffeine, and this is reflected in the more accurate predictions of aqueous solubility, with all four being within the experimental error of ± 0.6 log unit, and the mean of all four predictions being – 2.94.

It is recommended that, wherever possible, predictions be obtained from more than one software program and/or QSPR, and that the mean of all the predictions be used, unless one of the predicted values is clearly very different from the others, in which case that prediction should be rejected.

Abshear et al [2006] have demonstrated, using aqueous solubility prediction, that a consensus model can give much better predictions than individual predictive models.

2.9 Predictions using artificial neural networks (ANNs)

Taskinen and Yliruusi [2003] have discussed the prediction of physicochemical properties using neural network modelling. Clearly, an ANN is not transparent, and there is a risk of overtraining it, so it has to be used carefully, by someone who knows what they are doing. On the other hand, commercial software that uses the ANN approach has already been trained, so a non-expert user can use it quite safely. Since one cannot always expect all the OECD principles to be satisfied before using a QSPR or software program, lack of transparency is not a huge barrier to the use of an ANN method.

2.10 Potential pitfalls in the use of QSPRs

1. The compound of interest should be within the applicability domain of the QSPR/software program. This is generally not easy to determine. Most, but not all, software developers make their training sets available, but even then it is not always obvious whether one's compound of interest is within the applicability domain, because software developers do not provide tables of descriptors from which one could check applicability. However, the read-across approach can be used here. That is, one can use the software or QSPR to make predictions for similar compounds whose property values are known. If those predictions are acceptable, then it is reasonable to assume that the prediction for one's compound of interest will also be acceptable. Almost invariably, QSPRs and property prediction software are trained on organic compounds, and cannot handle inorganic compounds or metallo-organics (an exception to this is the SPARC software).

2. The user must be clear as to which endpoint is being predicted. This is particularly important when a software program is able to predict a number of similar endpoints. For example, several commercially available software programs for the prediction of aqueous solubility offer several endpoints, such

as solubility in pure water and intrinsic solubility (i.e. the solubility of the undissociated species).

3. It is essential to check that the units of the property being predicted are known and understood. For example, a predicted log (solubility) value will probably have solubility in moles per litre, whereas the user might think that it is in milligrammes per 100 ml.

4. Walker and de Wolf [2003] have warned against using a predicted property to predict another property. However, sometimes this is unavoidable, e.g. in the case where a compound has not been synthesised. In such cases one must accept that the accuracy of prediction will probably be lower than would otherwise be the case.

5. When using a QSPR, it is essential to check that it has been validated, preferably by use of external validation, or, failing that, by cross-validation; this is because it is possible to develop a QSPR that models the training set data well, but does not give good predictions. In the case of external validation, the prediction errors of the test set should be similar to those of the training set. In the case of cross-validation, the cross-validated R^2 (Q^2) value should not be < 0.5 [Eriksson et al 2003], and should not be more than 0.3 lower than the R^2 value [Walker et al 2003].

6. The calculation of descriptors for use in a QSPR should always be done using the same software as that used by the workers who developed the QSPR. The reason for this is that different software programs can yield different numerical values for a given descriptor; this is especially so for quantum chemical descriptors.

2.11 Major sources of misinterpretation of QSPR endpoints

1. Selection of wrong endpoint: e.g. intrinsic solubility instead of solubility in pure water.

2. Use of incorrect units for a property: e.g. g/100 ml instead of mol L^{-1}; use of natural logarithm instead of logarithm to base 10.
3. Use of a QSPR or software program to make predictions outside its applicability domain.
4. Placing too much reliance on a single prediction.

2.12 Criteria for good and rigorous read-across

The reliability of a QSPR property estimation can be judged by predicting values of the same property for one or more similar chemicals for which measured values of the property are available; if such predictions are judged to be acceptable (i.e. within or close to the experimental error on the measured property), then it can be assumed that the prediction for the chemical of interest will also be acceptable. However, this assumption depends on how similar to the chemical of interest are the "similar" chemicals [Barratt 2003]. Sedykh and Klopman [2006] have recently published an interesting read-across approach to the prediction of log K_{ow}, which they claim is superior to conventional group contribution methods.

Strictly, one ought to perform a statistical similarity exercise in order to obtain a numerical indication of similarity. However, this is clearly beyond the scope of a chemist working in a chemical company, and who probably has little or no knowledge of QSPR or similarity assessment. It is therefore suggested that a visual assessment of similarity should suffice. For example, if 4-chlorophenol were the chemical of interest, then suitable similar chemicals could be 3-chlorophenol, 4-bromophenol, or 3,5-dichlorophenol, but not, say, 4-chloronitrobenzene. 4-nonylphenol or pentachlorophenol.

3. Prediction of octanol-water partition coefficient (log K_{ow}, log P)

The octanol-water partition coefficient is the ratio of concentrations of a chemical in *n*-octanol and in water at equilibrium at a specified temperature (typically 25°C, although partition coefficient is not usually very temperature-dependent [Dearden & Bresnen 2005]). In the case of ionisable chemicals, it relates only to the concentration

ratio of the unionised species. Partition coefficient is a good surrogate for partitioning of chemicals through lipid membranes, and is the most important physicochemical descriptor of biological activity, appearing in some 70% of QSARs.

The term distribution coefficient (D_{ow}) is used for the ratio of total concentrations (both ionised and unionised species) in the two solvents, although it is generally assumed, not entirely correctly, that ionised species are not soluble in octanol. It should be remembered in this respect that water-saturated octanol contains about 27% mole of water at room temperature. For an ionisable chemical, it is possible to calculate log D_{ow} at a given pH from the value of log K_{ow}, given the pKa of the chemical:

$$\log D_{ow} = \log K_{ow} - \log (1 + 10^{A(pH - pKa)}) \qquad (3)$$

where A = 1 for acids and –1 for bases.

Hence those software packages that can calculate both log K_{ow} and pKa also offer log D_{ow} calculation.

Many publications have dealt with the estimation of log K_{ow} values from molecular structure, and Lyman [1990], Schwarzenbach et al [1993], Nendza [1998], Reinhard and Drefahl [1999], Leo [2000], Mannhold and van de Waterbeemd [2001], Livingstone [2003] and Klopman and Zhu [2005] have reviewed prediction methods for log K_{ow}; Livingstone [2003] in particular gives a detailed critical analysis of available methods. The main prediction methodologies are based on physicochemical, structural and/or topological descriptors, or on atomic or group contributions.

The earliest work on log K_{ow} prediction was that of Hansch and co-workers, who developed [Fujita et al 1964] a hydrophobic substituent constant π, which was, to a first approximation, additive, although it required numerous correction factors. Rekker and co-workers [Nys & Rekker 1973, Rekker 1977] developed a fragmental approach which proved easier to use. Extension of the fragmental approach by Leo et al [1975] led to the development of the ClogP software [www.daylight.com] for log K_{ow} prediction.

Bodor et al [1989] developed a QSPR with 14 physicochemical and quantum chemical descriptors to model log K_{ow} of a diverse set of 118 organic chemicals, with $R^2 = 0.882$ and a standard error of 0.296 log unit. The method of Ghose et al [1988] used atomic contributions, and on a set of 893 compounds the standard error was 0.496 log unit. Klopman and Wang [1991] used their MCASE group contribution approach to predict the log K_{ow} values of 935 organic compounds with a standard error of 0.39 log unit. This error is close to the experimental error on log K_{ow}.

A method of predicting log K_{ow} values that provides mechanistic insight is that of Abraham et al [1994b]. Using their solvatochromic descriptors they developed the following QSPR:

$$\log K_{ow} = 0.088 + 0.562\,R - 1.054\,\pi^H + 0.034\,\Sigma\alpha^H - 3.460\,\Sigma\beta^H + 3.814\,V_x \qquad (4)$$

$$n = 613 \quad R^2 = 0.995 \quad s = 0.116$$

where R = excess molar refractivity, π^H = a polarity term, $\Sigma\alpha^H$ and $\Sigma\beta^H$ = hydrogen bond donor and acceptor abilities respectively, and V_x = the McGowan characteristic molecular volume. Since the descriptors are approximately autoscaled, the magnitudes of the coefficients give an indication of the relative contribution of each descriptor to log K_{ow}. Thus it can be seen that hydrogen bond acceptor ability and molecular size make the most important contributions to log K_{ow}; on the other hand the contribution of hydrogen bond donor ability is negligible, and this is attributed to the hydrogen bond acceptor abilities of both water and octanol being very similar, while in contrast the hydrogen bond donor ability of water is very strong, accounting for the high negative coefficient on the $\Sigma\beta^H$ term. The standard error is very low, and may indicate some over-fitting of the data.

Although measured values of the Abraham descriptors are not available for all compounds, they can be calculated using the Absolv-2 software.

It is important to point out that, contrary to what might be thought, solubility in octanol is not a measure of lipophilicity. When a chemical is taken up by lipid *in vivo*,

it is always from an aqueous phase, and so it is the distribution between aqueous and lipid phases that is important, and not the absolute solubility in lipid. In fact, the term hydrophobicity is preferable to lipophilicity, because the driving force for transfer from water to lipid comes largely from the aqueous phase; that is, a chemical is pushed from water to lipid, rather than being pulled by lipid from water. The driving force has a large entropic component [Dearden & Bresnen 2005] because of water-structuring. Octanol tends to behave much as an ideal solvent, and solubility in octanol (S_o) is inversely correlated with melting point, but not with octanol-water partition coefficient. Dearden [1990] showed that the correlation between log K_{ow} and log S_o is very poor (n = 35, r^2 = 0.216, s = 0.512).

It is also pointed out that the calculation of log K_{ow} from the ratio of solubilities in octanol and water is rather inaccurate, as the results below show [Yalkowsky et al 1983]:

Table 4. Performance of log (S_o/S_w) compared with measured log K_{ow}

Solute	log(S_o/S_w)	log K_{ow}
Antipyrine	-0.73	0.26
Ethyl 4-aminobenzoate	1.86	1.96
Caffeine	-0.75	-0.20
Theophylline	-0.57	-0.09

Other workers [Miller et al 1985, Anliker & Moser 1987, Niimi 1991, Sijm et al 1999] have also shown that log (S_o/S_w) is not a good surrogate for log K_{ow}.

There are numerous software programs available for the estimation of log K_{ow} of organic chemicals, and some of these give good predictions. A recent comparison of 14 such programs [Dearden et al 2003a] found that, using a 138-chemical test set, the percentage of chemicals with log K_{ow} predicted within ± 0.5 log unit of the measured log K_{ow} value ranged from 94% to 50 %. The performances of the top six programs are shown in Table 5.

Table 5. Performance of top software for prediction of log K_{ow}

Software	% Predicted within ± 0.5 log unit of measured value	Standard error (log unit)
ADMET Predictor	94.2%	0.27
ACD/Labs	93.5%	0.27
ChemSilico	93.5%	0.30
Episuite	89.1%	0.34
SPARC	88.5%	0.33
ClogP	88.4%	0.29

Software programs not tested by Dearden et al [2003a] are Admensa, ASTER, AUTOLOGP [Devillers et al 1995], ChemAxon, ChemOffice, ChemProp, Molecular Modeling Pro, Pipeline Pilot and VCCLAB. Admensa is reported to yield a test set RMS error of 0.44 log unit. AUTOLOGP is reported [Devillers et al 1995] to yield a standard error of 0.39 log unit for a heterogeneous set of 800 organic compounds; it is not clear whether AUTOLOGP is still available. ChemOffice is reported to yield a standard error of 0.43 log unit, but that rises to 0.83 log unit for compounds possessing intramolecular hydrogen bonding. VCCLAB is reported to yield a standard error of 0.26 log unit. The performance of the other software programs is not known.

Sakuratani et al [2007] tested six software programs, using a test set of 134 simple organic compounds. None of the programs predicted log K_{ow} values of all the compounds. Their results were: Episuite (KOWWIN), n = 130, s = 0.94; ClogP, n = 131, s = 0.95; ACD/Labs, n = 127, s = 1.09; VLOGP [www.accelrys.com], n = 122, s = 1.11; SLOGP [www.chemcomp.com], n = 132, s = 1.34; COSMO [www.cosmologic.de], n = 129, s = 1.35.

It is recommended that at least two of the better software programs mentioned above be used for the prediction of log K_{ow}. If possible, the average of several predictions should be taken. It should be noted that the VCCLAB web-site (see Table 1), as well as giving its own log K_{ow} prediction, gives predictions from six other software packages, together with the mean of all seven.

4. Prediction of aqueous solubility

Aqueous solubility depends not only on the affinity of a solute for water, but also on its affinity for its own crystal structure. Molecules that are strongly bound in their crystal lattice require considerable energy to remove them. This also means that such compounds have high melting points, and in general high-melting compounds have poor solubility in any solvent.

Removal of a molecule from its crystal lattice means an increase in entropy, and this can be difficult to model accurately. For this reason, as well as the fact that the experimental error on solubility measurements can be quite high (generally reckoned to be about 0.6 log unit), the prediction of aqueous solubility is not as accurate as is the prediction of partition coefficient. Nevertheless, many papers [Dearden 2006] and a book [Yalkowsky & Banerjee 1992] have been published on the prediction of aqueous solubility, as well as a number of reviews [Lyman 1990, ECETOC 1998, Reinhard & Drefahl 1999, Mackay 2000, Schwarzenbach et al 2003, Dearden 2006]. There are also a number of commercial software programs available for that purpose [ECETOC 2003, Dearden 2006]. Livingstone [2003] has discussed the reliability of aqueous solubility predictions from both QSPRs and commercial software.

It should be noted that there are various ways that aqueous solubilities can be reported: in pure water, at a specified pH, at a specified ionic strength, as the undissociated species (intrinsic solubility), or in the presence of other solvents or solutes. Solubilities are also reported in different units, for example g/100 ml, mol L^{-1}, mole fraction. The use of mol L^{-1} is recommended, as this provides a good basis for comparison.

Hansch et al [1968] first reported the inverse correlation between the aqueous solubility (S_{aq}) of liquids and their octanol-water partition coefficients (K_{ow});

$$\log S_{aq} = -1.339 \log K_{ow} + 0.978 \qquad (5)$$

$$n = 156 \quad r^2 = 0.874 \quad s = 0.472$$

Lyman [1990] lists 18 (log S_{aq} *vs.* log K_{ow}) QSPRs for various classes of chemicals. So, for example, for a liquid ketone one could use the QSPR for ketones developed by Hansch et al [1968]:

$$\log S_{aq} = - 1.229 \log K_{ow} + 0.720 \qquad (6)$$

$n = 13 \quad r^2 = 0.960$

The log K_{ow} value could be either a measured or a calculated value (see section 3 on octanol-water partition coefficient).

However, for solids work has to be done to remove molecules from their crystal lattice, and the simplest way to account for this is to use what Yalkowsky and co-workers have termed the General Solubility Equation (GSE), which incorporates a melting point term to account for the behaviour of solids [Sanghvi et al 2003]:

$$\log S_{aq} = 0.5 - \log K_{ow} - 0.01(MP - 25) \qquad (7)$$

where MP is the melting point (^oC). The melting point term is taken as zero for compounds melting at or below 25^oC. Calculated log K_{ow} and MP values can be used in the GSE, although measured values are preferred. Aqueous solubilities of 1026 non-electrolytes, with a log S_{aq} range of – 13 to + 1 (S in mole L^{-1}), calculated with the GSE had a standard error of 0.38 log unit.

Yalkowsky and co-workers have also developed the AQUAFAC group contribution method for calculating aqueous solubility [Myrdal et al 1995]. They calculated the ideal solubility or fugacity ratio F as:

$$\log F = - (56.5 - 19.2 \log \sigma)(MP - 25)/5706 \qquad (8)$$

where σ = a symmetry number, i.e. the number of indistinguishable positions in which a molecule can be oriented, and the units in the equation are SI units.

For liquids, log S_{aq} = - log γ_m, and for solids log S_{aq} = log F – log γ_m, where γ_m is the molar activity coefficient, which itself is given by:

$$\log \gamma_m = \Sigma n_i q_i \qquad (9)$$

where n_i is the number of times a group appears in a molecule and q_i is the contribution of that group. Mackay [2000] lists a large number of the group contribution values. For a set of 97 diverse chemicals, the AQUAFAC mean absolute error of prediction was 0.41 log unit, whilst that using the log K_{ow} approach was 0.61 log unit. As usual, there is a trade-off between accuracy and ease of use.

Good predictions for a large diverse data set have been obtained by the use of linear solvation energy descriptors [Abraham & Le 1999]:

$$\log S_{aq} = 0.518 - 1.004\, R + 0.771\, \pi^H + 2.168\, \Sigma\alpha^H + 4.238\, \Sigma\beta^H - 3.362\, \Sigma\alpha^H.\Sigma\beta^H - 3.987\, V_X \qquad (10)$$

n = 659 R^2 = 0.920 s = 0.557

where R = excess molar refractivity (a measure of polarisability), π^H = a polarity/polarisability term, $\Sigma\alpha^H$ and $\Sigma\beta^H$ = sums of hydrogen bond donor and acceptor abilities respectively, and V_X = McGowan characteristic molecular volume. All of these terms can be calculated with the Absolv-2 software (see Table 1). It can be seen from the Abraham and Le equation that the main factors controlling aqueous solubility are hydrogen bond acceptor ability and molecular size.

Katritzky et al [1998] used their CODESSA descriptors to model the aqueous solubilities of a large diverse set of organic chemicals:

$$\log S_{aq} = -16.1\, Q_{min} - 0.113\, N_{el} + 2.55\, FHDSA(2) + 0.781\, ABO(N) + 0.328\, {}^0SIC - 0.0143\, RNCS - 0.882 \qquad (11)$$

n = 411 R^2 = 0.879 s = 0.573

where Q_{min} = most negative partial charge, N_{el} = number of electrons, FHDSA(2) = fractional hydrogen bond donor area, ABO(N) = average bond order of nitrogen atoms, ^{0}SIC = an information content topological descriptor, and RNCS = relative negatively charged surface area. The CODESSA software is available from SemiChem Inc. [www.semichem.com].

Electrotopological state descriptors [Votano et al 2004], hydrogen bonding and nearest-neighbour similarities [Raevsky et al 2004] and group contributions [Klopman & Zhu 2001] have also been used to model the aqueous solubilities of large diverse data sets of organic chemicals.

There are relatively few studies of solubility prediction within specific chemical classes. Hawker and Connell [1988] obtained the following QSPR for polychlorinated biphenyls (PCBs) with 1-10 chlorine atoms:

$$\log S_{aq} = (-4.13 \times 10^{-2})\ TSA + (23.8/R)(1 - T_m/T) + 3.48 \qquad (12)$$

$n = 17 \quad R^2 = 0.901 \quad s = 0.464$

where TSA = total surface area, R = universal gas constant, T_m = melting point (K) and T = temperature at which solubility is required (K).

Huuskonen et al [1997] used artificial neural network modelling to predict the aqueous solubilities of steroids and other drug classes. For a set of 28 steroids, with a log S_{aq} range of – 5.4 to – 2.6 (S in mole L^{-1}), they obtained a standard error of 0.29 log unit, using 5 molecular connectivity descriptors.

Yang et al [2007] found a good correlation of log S_{aq} with mean molecular polarisability for a small set of dioxins ($n = 12$, $r^2 = 0.978$, $s = 0.30$).

Solubility can vary considerably with temperature, and it is important that solubility data are reported at a given temperature.

Dearden et al [2003b] compared 11 commercial software programs for aqueous solubility prediction (as log S), and found considerable variation in performance

against a 113-chemical test set of organic chemicals that included 17 drugs and pesticides. The performance of the top four programs is shown in Table 6.

Table 6. Performance of top software for prediction of log Saq for 113-compound test set

Software	% Predicted within ± 0.5 log unit of measured value	Standard error (log unit)
ChemSilico	75.0%	0.49
ADMET Predictor	74.3%	0.50
ACD/Labs	72.6%	0.50
Episuite (WSKOWWIN)	69.9%	0.56

The Episuite predictions were made without the input of measured melting point values. Dearden [2007] tested the new fragment-based WATERNT module in the Episuite software on the same 113-compound test set, and found it to be better than all previously tested software (79.6% within ± 0.5 log unit of measured value; standard error = 0.44 log unit).

Dearden [2006] tested 16 commercially available software programs for their ability to predict the aqueous solubility of a 122-compound test set of drugs with accurately measured solubilities in pure water. Again there was considerable variation in performance. The performance of the top five programs is shown in Table 7.

Dearden [2007] tested the new fragment-based WATERNT module in the Episuite software on the same 122-drug test set, and found it to be among the worst of all previously tested software (38.5% within ± 0.5 log unit of measured value; standard error = 0.93 log unit). Investigation indicated that this was caused by the program's not including all fragments and/or correction factors in its calculations. Software not tested by Dearden et al [2003b] or Dearden [2006] are ASTER, ChemProp, Molecular Modeling Pro and Pipeline Pilot. Their performances are not known.

Table 7. Performance of top software for prediction of log Saq for 122-drug test set

Software	% Predicted within ± 0.5 log unit of measured value	Standard error (log unit)
Admensa	72.1%	0.65
ADMET Predictor	64.8%	0.47
ChemSilico	59.8%	0.73
ADME Boxes	59.0%	0.62
ACD/Labs	59.0%	0.66

It is recommended that at least one of the above software programs be used for the prediction of aqueous solubility as log S_{aq}. If possible, the average of several predictions should be taken (see Table 3).

5. Prediction of pKa

Within a congeneric series of chemicals, pKa is often closely correlated with the Hammett substituent constant, and this is the basis for a number of attempts at pKa prediction. Harris and Hayes [1990] and Livingstone [2003] have reviewed the published literature in this area.

The Hammett substituent constant σ was derived from a consideration of acid dissociation constants Ka, and most non-computerised methods of calculating Ka and pKa values are based on σ values:

$$\text{pKa (derivative)} = \text{pKa (parent)} - \rho\sigma \qquad (13)$$

where ρ is the series constant, which is 1.0 for benzoic acids. Harris and Hayes [1990] list ρ values for other series.

Harris and Hayes [1990] give several examples of pKa calculation, for example for 4-t-butylbenzoic acid. The pKa value of benzoic acid is 4.205, the ρ value for benzoic acids is 1.0, and the σ value for 4-t-butyl is – 0.197. Hence the pKa value of 4-t-

butylbenzoic acid is calculated as 4.205 – (– 0.197) = 4.402. This value is virtually identical to the measured value for this compound.

A number of publications have dealt with estimation of pKa values from chemical structure, but these relate mostly to specific chemical classes, e.g. amines [Nagy et al 1989], 4-aminoquinolines [Kaschula et al 2002] and imidazol-1-ylalkanoic acids [Soriano et al 2004]. There have, however, been a few attempts to model pKa values of diverse sets of chemicals. Klopman and Fercu [1994] used their MCASE methodology to model the pKa values of a set of 2464 organic acids, and obtained good predictions; a test set of about 600 organic acids yielded a standard error of 0.5 pKa unit. Klamt et al [2003] employed their COSMO-RS methodology to predict pKa values of 64 organic and inorganic acids, with a standard error of 0.49 pKa unit.

There are a number of software programs that predict multiple pKa values of organic chemicals, but there are no published comparisons of their performance, although a comparison has recently been published of the performance of four software packages for the calculation of pKa values of ionisable groups in proteins [Davies et al 2006]. ACD/Labs has a claimed standard error of 0.39 pKa unit for 22 compounds, and one of 0.36 pKa unit for 26 drugs. pKalc (part of the PALLAS suite) is claimed to be accurate to within 0.25 pKa unit [Tsantili-Kakoulidou et al 1997], QikProp is claimed to have a mean absolute error (MAE) of 0.19 pKa unit, and SPARC is claimed to have a RMS error of 0.37 pKa unit when evaluated on 3685 compounds[Hilal & Karickhoff 1995]. ADMET Predictor is claimed to have a MAE of 0.56 pKa unit for a test set of 2143 diverse chemicals. ChemSilico is reported to have a MAE of 0.99 pKa unit for a test set of 665 diverse chemicals, many of them multiprotic. No published information appears to be available for the performance of ADME Boxes, ASTER, Pipeline Pilot and VCCLAB.

Dearden and Lappin [2007] have tested the performance of the ten available software programs that calculate pKa values (see Table 2; ASTER is not currently available). Some of these programs will calculate pKa values of all ionisable sites. However, the test-set of 665 chemicals that they used, which was kindly supplied by ChemSilico Inc. and used by them as their test-set, had measured pKa values only for the prime ionisation site in each molecule. There were doubts about the correct structures of 11

of the test set chemicals, and so the programs were tested on 654 chemicals. Some of the software companies kindly ran our compounds through their software in-house, and we have not yet obtained clearance from them to release their results. The results given in Table 8 are for those software programs that we have in our laboratory, or which are freely available on-line, together with the best-performing software (ADME Boxes from Ap-Algorithms).

Table 8. Performance of pKa prediction software using a test set of 654 diverse organic chemicals

Software	Number of chemicals handled	r^2	MAE
ADME Boxes	628	0.959	0.32
VCCLabs	611	0.931	0.40
SPARC	645	0.848	0.78
ChemSilico			0.99*
ACD/Labs	645	0.682	1.07
Pallas	647	0.661	1.17

*MAE reported on the ChemSilico website

It should be noted that some of our test-set chemicals could have been included in the training sets for the various pKa prediction software programs. The only software where that was not the case is, of course, ChemSilico. This is therefore one reason for the ChemSilico predictions being rather poor relative to some of the others.

It is recommended that the VCCLabs and SPARC software, which are both free to use on-line, be used for pKa prediction. It should be noted, however, that the VCCLabs program failed to predict pKa values for 43 compounds in our 654-compound test-set.

6. Prediction of melting point

Melting point is an important property for two main reasons. Firstly, it indicates whether a chemical will be solid or liquid at particular temperatures, which will dictate how it is handled. Secondly, it is used in the General Solubility Equation [Sanghvi et al 2003] to predict aqueous solubility.

The melting point of a crystalline compound is controlled largely by two factors – intermolecular interactions and molecular symmetry. For example, 3-nitrophenol, which can hydrogen-bond via its –OH group, melts at 97^{o}C, whereas its methyl derivative, 3-nitroanisole, which cannot hydrogen-bond with itself, melts at 39^{o}C. The symmetrical 1,4-dichlorobenzene melts at 53^{o}C, whilst the non-symmetrical 1,3-dichlorobenzene melts at -25^{o}C. These and other effects have been discussed in detail by Dearden [1999].

There have been many attempts to predict the melting point of organic chemicals, and these have been reviewed by Horvath [1992], Reinhard and Drefahl [1999], Dearden [1999, 2003] and Tesconi and Yalkowsky [2000]. It may be noted that in the 19th century Mills [1884] developed a QSPR based on carbon chain length for melting points of homologous series of compounds that was accurate to ± 2^{o}.

Essentially two approaches have been used in the prediction of melting point – the physicochemical/structural descriptor approach and the group contribution approach. The former is exemplified by the work of Katritzky et al [1997], who used 9 of their CODESSA descriptors to model a diverse set of 443 aromatic chemicals with R^2 = 0.837 and s = 30.2^{o}. The CODESSA software is available from SemiChem Inc. [www.semichem.com]. This is a complex QSPR, with descriptors that are not easy to comprehend, and reflects the difficulty of modelling the melting points of diverse data sets. Even for a set of 58 PCB congeners with 1-10 chlorine atoms, a 5-term QSPR was required [Abramowitz & Yalkowsky 1990], with R^2 = 0.83 and s = 22.1^{o}.

Yalkowsky and co-workers have published extensively on the prediction of melting point. They incorporated terms to account for conformational flexibility and rotational

symmetry [Tsakanikas & Yalkowsky 1988] and molecular eccentricity [Abramowitz & Yalkowsky 1990] to try to account for the entropic contributions to melting. They were able [Zhao & Yalkowsky 1999] to model the melting points of 1040 aliphatic chemicals, using a combination of molecular geometry and group contributions, with a standard error of 34.4°.

Todeschini et al [1997] used their WHIM descriptors to model the melting points of 94 European Union environmental priority chemicals, with a standard error of 32.8°. Bergström et al [2003] used principal components analysis and partial least squares to model the melting points of 227 diverse drugs. They used 2-D, 3-D and a combination of 2-D and 3-D descriptors to give three separate models. A consensus of all three models gave the best results, with R^2 = 0.63 and RMS error = 35.1°. Modarresi et al [2006] used eight descriptors from Tsar [www.accelrys.com], CODESSA [www.semichem.com] and Dragon [www.virtuallaboratory.org/lab/edragon] to model the melting points of 323 drugs, with R^2 = 0.660 and RMS error = 41.1°.

Recently Karthikeyan et al [2005] used a very large diverse training set of 4173 chemicals to develop a QSPR based on a neural network approach using principal components. They found 2-D descriptors to be better than 3-D descriptors; the results were as follows:

Training set	Internal validation	Test set	Test set (drugs)
(n = 2089)	(n = 1042)	(n = 1042)	(n = 277)
R^2 = 0.661	Q^2 = 0.645	Q^2 = 0.658	Q^2 = 0.662
MAE = 37.6°	MAE = 39.8°	MAE = 38.2°	MAE = 32.6°

Considering the size and diversity of the data sets, the statistics are quite good. However, the methodology used was complex, and could not readily be applied.

The group contribution approach to melting point prediction was first used by Joback and Reid [1987]. Simamora and Yalkowsky [1994] modelled the melting points of a diverse set of 1690 aromatic compounds using a total of 41 group contributions and four intramolecular hydrogen bonding terms, and found a standard error of 37.5°.

Constantinou and Gani [1994] used two levels of group contributions to model the melting points of 312 diverse chemicals, and obtained a mean absolute error (MAE) of prediction of 14.0°, compared with a MAE of 22.6° for the Joback and Reid method. Marrero and Gani [2001] extended this approach to predict the melting points of 1103 diverse chemicals with a standard error of 25.3°. Tu and Wu [1996] used group contributions to predict melting points of 1310 diverse chemicals with a MAE of 8.2%.

There are several software programs that predict melting point; they all use one or more group contribution approaches. Dearden [2003] used a 96-compound test set to compare the performances of three of these programs. Episuite calculates melting point by two methods, that of Joback and Reid [1987] and that of Gold and Ogle [1969], and takes their mean. ChemOffice uses the method of Joback and Reid [1987], and ProPred uses the Gani approach [Constantinou & Gani 1994, Marrero & Gani 2001].

Table 9. Software performance for prediction of melting point of a 96-compound test set

Software	Mean absolute error
Episuite	26.3°
ChemOffice	27.0°
ProPred	25.8°

An ECETOC [2003] report mentions a US. Environmental Protection Agency (EPA) report [1999] concerning the performance of the Episuite MPBPVP module; for two large, diverse test sets the performance was: (i) n = 666, r^2 = 0.73, MAE = 45°; (ii) n = 1379, r^2 = 0.71, MAE = 44°. The lower MAE values reported in Table 7 could reflect either less diversity in the 96-compound test set used by Dearden [2003], or improvements made in the software since 1999. Molecular Modeling Pro uses the Joback and Reid [1987] method, so its performance should be the same as that of ChemOffice. The performances of ChemProp and PREDICT are not known.

It can be seen that there is little to choose between the programs in terms of accuracy of prediction. They can all operate in batch mode. It is therefore recommended that

the Episuite software, which is freely downloadable, and at least one other method be used to calculate melting point.

It should be noted that currently both QSPR methods and software programs have prediction errors well in excess of the error on experimental measurement of melting point, which is usually < 2^o. Therefore it is preferable to use measured melting points if at all possible.

7. Prediction of boiling point

Boiling point (T_b) is an important property since it is an indicator of volatility, and can be used to predict vapour pressure. From the Clausius-Clapeyron equation, boiling point is inversely proportional to the logarithm of vapour pressure. Boiling point also indicates whether a chemical is gaseous or liquid at a given temperature.

Lyman [2000] has discussed seven recommended methods for the prediction of boiling point. The methods are based on physicochemical and structural properties and group contributions. Perhaps the simplest of those methods is that of Banks [1939], who developed the following QSPR:

$$\log T_b\ (K) = 2.98 - 4/\sqrt{MW} \qquad (14)$$

where MW = molecular weight. No statistics were given for this QSPR.

Rechsteiner [1990], Reinhard and Drefahl [1999] and Dearden [2003] have reviewed the QSPR prediction of boiling point.

Many studies of boiling point prediction have dealt with specific chemical classes, and very good correlations have generally been obtained. In 1884 Mills [1884] modelled the boiling points of a number of homologous series with QSPRs based on carbon chain length. Ivanciuc et al [2000] used 4 topological descriptors to model the boiling points of 134 alkanes with a standard error of 2.7^o, whilst Gironés et al [2000] used only one quantum chemical descriptor (electron-electron repulsion energy) to model the boiling points of 15 alcohols with a standard error of 5.6^o.

Models based on diverse training sets are, however, more widely applicable. Katritzky et al [1996a] used 4 CODESSA descriptors to model the boiling points of 298 diverse organic compounds:

$$T_b\ (K) = 67.4\ G_I^{1/3} + 21540\ HDSA(2) + 140.4\ \delta^-_{max} + 17.5\ N_{Cl} - 151.3 \qquad (15)$$

$n = 298 \quad R^2 = 0.973 \quad s = 12.4^o$

where G_I = gravitational index, HDSA(2) = area-weighted surface charge of hydrogen-bond donor atoms, δ^-_{max} = most negative atomic partial charge, and N_{Cl} = number of chlorine atoms. The CODESSA software is available from SemiChem Inc. (www.semichem.com).

Wessel and Jurs [1995] used their ADAPT descriptors to develop two QSPRs for the prediction of boiling point – one for compounds containing O, S and halogens, and the other for compounds containing N. The QSPR for O, S and halogens is:

$$T_b\ (K) = 0.3009\ PPSA - 3.690\ PNSA - 51.78\ RPCG + 9.515\ N_{RA} + 19.21\ SQMW + 554.7\ SADH - 25.52\ N_F + 19.52\ KETO + 50.84\ N_{sulf} - 135.0\ S/NA + 59.86 \qquad (16)$$

$n = 248 \quad R^2 = 0.991 \quad RMS\ error = 11.6^o$

where PPSA = partial positive surface area, PNSA = partial negative surface area, RPCG = relative positive charge, N_{RA} = number of ring atoms, SQMW = square root of molecular weight, SADH = surface area of donatable hydrogen atoms , N_F = number of fluorine atoms, KETO = indicator variable for ketone, N_{sulf} = number of sulphide groups, and S/NA = (number of sulphur atoms)/(total number of atoms). The ADAPT descriptors are available in the Pharma Algorithms ADME Boxes software.

Basak et al [2001] used 8 topochemical, topological and hydrogen bonding descriptors to model the boiling points of 1015 diverse organic compounds, with a standard error of 15.7°. The best QSPR developed to date is that of Hall and Story [1996], who used atom-type electrotopological descriptors [Kier & Hall 1999] and a

neural network to obtain a MAE of 3.9° for a set of 298 diverse chemicals with a boiling point range of about 430°.

The group contribution approach was used first by Joback and Reid [1987], who obtained a mean absolute error (MAE) of 12.9° for a set of 438 diverse chemicals. Stein and Brown [1994] devised a simple group contribution method to model boiling points of a very large set of 4426 diverse chemicals, with a MAE of 15.5°. A group contribution approach was also used by Marrero and Gani [2001] to model the boiling points of 1794 organic compounds with a standard error of 8.1°, whilst Labute [2000] used 18 atomic contributions on a set of 298 diverse organics, to give a standard error of 15.5°. Simamora and Yalkowsky [1994] used 36 group contributions and 4 intramolecular hydrogen bonding terms to model the boiling points of a diverse set of 44 aromatic compounds, with a standard error of 17.6°

There are a number of software programs available for the prediction of boiling point, and Dearden [2003] compared the performance of six of these using a 100-compound test set. The results were as follows:

Table 10. Software performance for prediction of boiling point of a 100-compound test set

Software	Mean absolute error
ACDLabs	1.0°
SPARC	6.3°
Episuite	13.8°
ChemOffice	13.8°
ProPred	16.1°
Molecular Modeling Pro	21.7°

The ACDLabs result is based on the 54 chemicals in the test set that were not included in the ACDLabs training set.

Clearly the ACDLabs software gives by far the best predictions, but has to be purchased. SPARC is freely accessible, but operates only in manual mode, with SMILES input. Episuite can be freely downloaded, but its standard error of prediction

is more than twice that of SPARC. ECETOC [2003] quotes the US. Environmental Protection Agency (EPA) [1999] testing of the MPBPVP module of the Episuite software; two very large diverse test sets yielded the following: n = 4426, MAE = 15.5°; n = 6584, MAE = 20.4°. These results are comparable with those given in Table 8 above. It is recommended that at least two predictions be obtained, and their average used.

Three other software programs, ASTER, ChemProp and PREDICT also predict boiling point. ASTER is claimed [ECETOC 1998] to have a mean error of 2%. PREDICT is claimed to have a MAE of 12.9°. The performance of ChemProp is not known.

8. Prediction of vapour pressure

The vapour pressure (VP) of a chemical controls its release into the atmosphere, and thus is an important factor in environmental distribution of chemicals. Vapour pressure is highly temperature-dependent. Most literature values are at ambient temperature, but some QSPRs allow predictions over a range of temperatures.

The variation of vapour pressure with temperature is given by the Clausius-Clapeyron equation:

$$\ln (VP_2/VP_1) = -(L/R)((1/T_2) - (1/T_1)) \quad (17)$$

where L = latent heat of vaporisation, and R = universal gas constant.

If the latent heat of vaporisation is high, vapour pressure changes markedly with temperature, which is why some chemicals (e.g. PCBs) deposit out in polar regions.

Numerous methods are available for the estimation of vapour pressure, and Grain [1990], Schwarzenbach et al [1993], Delle Site [1996], Sage and Sage [2000] and Dearden [2003] have reviewed many of these. The descriptors used in vapour pressure QSPRs include physicochemical, structural and topological descriptors, and group

contributions. Katritzky et al [1998] used their CODESSA descriptors to model the vapour pressure (in atmospheres at 25°C) of a large set of diverse organic chemicals:

$$\log VP = -0.00618\ G_I - 4.02\ HDCA(2) + 0.129\ SA\text{-}2(F) + 6.02\ MNAC(Cl) - 0.0143\ SA(N) + 2.30 \quad (18)$$

$n = 411 \quad R^2 = 0.949 \quad s = 0.331$

where G_I = gravitational index, HDCA(2) = hydrogen-bond donor charged surface area, SA-2(F) = sum of surface area of fluorine atoms, MNAC(Cl) = maximum net atomic charge for a chlorine atom, and SA(N) = sum of surface area of nitrogen atoms. The CODESSA software is available from SemiChem Inc. [www.semichem.com].

Liang and Gallagher [1998] used polarisability and 7 structural descriptors to model the vapour pressure of 479 diverse organic chemicals, using both multiple linear regression and an artificial neural network. There was little difference between the two methods with MLR giving a standard error of 0.534 log unit and ANN yielding 0.522 log unit.

Tu [1994] used a group contribution method to model the vapour pressure of 1410 diverse organic chemicals. Using 81 group contributions, 2 hydrogen bonding terms and melting point he obtained a standard error of 0.36 log unit.

The vapour pressures of 352 hydrocarbons and halohydrocarbons were modelled by Goll and Jurs [1999], using 7 of their ADAPT descriptors. Vapour pressure was recorded in pascals, and the data covered the log VP range – 1.016 to + 6.65.

$$\log VP = -0.670\ V0 + 0.204\ NF + 5.47 \times 10^{-2}\ NSB - 0.121\ NRA - 6.35 \times 10^{-2}\ DPSA + 0.117\ N3C + 0.518\ RPCG + 8.15 \quad (19)$$

$n = 352 \quad R^2 = 0.983 \quad$ RMS error = 0.186 log unit

where V0 = zero order molecular connectivity, NF = number of fluorine atoms, NSB = number of single bonds, NRA = number of atoms in ring systems, DPSA =

difference between partial positive surface area and partial negative surface area, N3C = number of 3rd order clusters, and RPCG = relative positive charge. The ADAPT descriptors are available in the ADME Boxes software [www.ap-algorithms.com].

Some of the ADAPT descriptors are difficult of interpretation, but have been found to give good correlations of a number of physicochemical properties. The very low standard error reflects the fact that there is little chemical diversity within the compounds used.

A number of studies [Andreev et al 1994, Kühne et al 1997, Yaffe & Cohen 2001] allow of the estimation of vapour pressures over a range of temperatures.

There are several commercially available software programs that will calculate vapour pressure; one of them (ACDLabs) will allow the calculation of vapour pressure over a temperature range. Using a 100-compound test set of organic chemicals with vapour pressures measured at 25°C, Dearden [2003] compared the performance of four software programs that calculate log (vapour pressure). The test results are given below.

Table 11. Software performance for prediction of vapour pressure of a 100-compound test set

Software	Mean absolute error *(log unit)*
SPARC	0.105
ACDLabs	0.107
Episuite	0.285
Molecular Modeling Pro	0.573

The programs can operate in batch mode, except for SPARC. The ACDLabs result was determined on only 42 compounds; 46 test set compounds that were used in the ACD/Labs training set were deleted, and in addition the ACD/Labs software did not give a vapour pressure at 25°C for 18 very volatile compounds. ECETOC [2003] quotes the US. Environmental Protection Agency (EPA.) [1999] testing of the

MPBPVP module of the Episuite software; n = 805, r^2 = 0.941, MAE = 0.476 log unit. This MAE probably reflects either the greater diversity of the US EPA. test set or improvements made in the software since 1999. Other software programs that calculate vapour pressure, but were not tested by Dearden [2003], are Absolv-2, ASTER, ChemProp, PREDICT and ProPred The prediction errors of the PREDICT software are reported to be 2 – 5%, depending on the method of calculation. The performances of the other software programs are not known although ASTER is expected [ECETOC 2003] to give similar results to those from the Episuite software.

It is recommended that either SPARC, Episuite or ACDLabs software be used for the calculation of vapour pressure. Predictions from at least two different sources should be obtained if possible.

9. Prediction of Henry's law constant (air-water partition coefficient)

The air-water partition coefficient is important in the distribution of chemicals between the atmosphere and water in the environment. The prediction of Henry's law constant (*H*) has been reviewed by Schwarzenbach et al [1993], Reinhard and Drefahl [1999], Mackay et al [2000] and Dearden and Schüürmann [2003].

One simple way of calculating *H* is to use the ratio of vapour pressure and aqueous solubility (vp/c_w). It is not a highly accurate method, but neither is the measurement of *H*, especially for chemicals with very high or very low *H* values. vp/c_w can be converted to the dimensionless form of *H* (ratio of concentrations in air and water, c_a/c_w, or K_{aw}) by the following equation, which is valid for 25°C:

$$c_a/c_w = 40.874\ vp/c_w \qquad (20)$$

Most prediction methods for *H* use a group or bond contribution approach, although some have used physicochemical properties [Dearden et al 2000]. The group and bond contribution methods were first used by Hine and Mookerjee [1974], who obtained, for a set of 263 diverse simple organic chemicals, a standard deviation of 0.41 log unit

for the group contribution method and one of 0.42 for the bond contribution method. Cabani et al. [1981] claimed an improvement in the group contribution method over that of Hine and Mookerjee, whilst Meylan and Howard [1991] extended the bond contribution method and obtained, for a set of 345 diverse chemicals, a standard error of 0.34 log unit. Their method, together with a group contribution method, is incorporated in the HENRYWIN module of the Episuite software.

Several workers have used physicochemical and/or structural descriptors to model *H*. Nirmalakhandan and Speece [1988] developed a QSPR using a polarisability descriptor, a molecular connectivity term and an indicator variable for hydrogen bonding. However, Schüürmann and Rothenbacher [1992] found it to have poor predictive power.

Russell et al [1992] used their ADAPT software to develop a 5-descriptor model of log K_{aw} for a relatively small but diverse data-set:

$$\log K_{aw} = -0.547\ N\text{HEAVY} + 0.0402\ \text{WPSA} + 0.0360\ \text{RNCS} + 10.1\ \text{QHET} - 215\ \text{QRELSQ} + 0.73 \qquad (21)$$

$n = 63 \quad R^2 = 0.956 \quad s = 0.375$

where *N*HEAVY = number of heavy atoms, WPSA = (total solvent-accessible surface area) x (sum of surface areas of positively charged atoms), RNCS = (charge on most negative atom) x (surface area of most negative atom)/(sum of charges on negatively charged atoms), QHET = (total charge on heteroatoms)/(number of heteroatoms), and QRELSQ = square of (total charge on heteroatoms)/(number of atoms). Note that the ADAPT descriptors are available in the Pharma Algorithms ADME Boxes software (www.ap-algorithms.com).

The Ostwald solubility coefficient *L* (the reciprocal of K_{aw}) of a very diverse data-set of chemicals was modelled by Abraham et al [1994a]:

$$\log L = 0.577\ R + 2.549\ \pi + 3.813\ \Sigma\alpha + 4.841\ \Sigma\beta - 0.869\ V_X + 0.994 \qquad (22)$$

$n = 408 \quad R^2 = 0.996 \quad s = 0.151$

where R = excess molar refractivity (a measure of polarisability), π = a polarity/polarisability term, α and β = hydrogen bond donor and acceptor abilities respectively, and V_X = the McGowan characteristic volume (see next section on prediction of relative density of liquids). The Abraham descriptors are approximately auto-scaled, so that the magnitudes of the coefficients in equation 22 indicate the relative contributions of each term. It is clear that hydrogen bonding is the most important factor controlling water-air distribution; the greater magnitude of the $\Sigma\beta$ term probably reflects the strong hydrogen bond donor ability of water. Molecular size, represented by V_X, appears to play only a minor role in determining air-water partitioning. It may be noted that the very high correlation coefficient and low standard error of equation 22 suggest possible overfitting; no external validation of equation 22 was provided. The Abraham descriptors are available in the Absolv-2 software (www.ap-algorithms.com).

Katritzky et al [1996b] used their CODESSA software [www.semichem.com] to model the data-set of Abraham et al [1994]:

$$\log L = 42.37\ \mathrm{HDCA}(2) + 0.65\ [N(\mathrm{O}) + N(\mathrm{N})] - 0.16\ \Delta\mathrm{E} + 0.12\ \mathrm{PCWT} + 0.82\ N_R + 2.65 \qquad (23)$$

$$n = 406 \quad R^2 = 0.942 \quad s = 0.52$$

where HDCA(2) = hydrogen bond donor ability, $N(\mathrm{O}) + N(\mathrm{N})$ = a linear combination of the number of oxygen and nitrogen atoms, ΔE = HOMO-LUMO energy difference, PCWT = most negative partial charge-weighted topological electronic index, and N_R = number of rings. It may be noted that the standard error of 0.52 log unit is more realistic than is that of 0.151 reported by Abraham et al [1994].

Katritzky et al [1998] used predicted vapour pressure and aqueous solubility to calculate Henry's law constant according to equation 20 for 411 diverse chemicals. The table giving their results was inadvertently omitted in their paper, but they reported a standard error of 0.63 log unit, which is not very much greater than that found (0.52 log unit) in their correlation shown in equation 23 above.

Very recently QSPRs have been developed by Modarresi et al [2007] using a very large (940-compound) diverse data set. Using genetic algorithm selection of descriptors, they obtained a 10-descriptor QSPR with a root mean square error of 0.571 log unit.

There are seven software programs that calculate Henry's law constant, namely Episuite, Absolv-2, ADME Boxes, ASTER, ChemProp, ProPred and SPARC. The performances of the last five are not known.

Dearden and Schüürmann [2003] tested a number of methods for prediction of log *H*, using a large, diverse test set of 700 chemicals. Only one of the methods, the bond contribution method in the HENRYWIN module of the Episuite software, allowed prediction of log *H* for all 700 chemicals, with a mean absolute error of prediction of 0.63 log unit.

It is recommended that the HENRYWIN module of the Episuite software be used for the prediction of Henry's law constant.

10. Prediction of relative density of liquids

Nelken [1990] and Reinhard and Drefahl [1999] have reviewed the prediction of relative density, ρ_L. A related property is molar volume, V_M (the volume in cm^3 occupied by 1 gram mole of a compound), and the two are related thus:

$$\rho_L = MW/V_M \qquad (24)$$

where MW = molecular weight, and ρ_L has the units of g cm^{-3}.

Correlations between density or molar volume and molecular surface area [Grigoras 1990], molecular connectivites [Kier & Hall 1976] and group contributions [Girolami 1994] have been reported. The Girolami method is very simple, and is based on the following equation:

$$\rho_L = MW(5V_{scal})^{-1} \quad (25)$$

where M = molecular weight, and V_{scal} = scaled volume calculated as the sum of the atom contributions of the constituent atoms. The method is claimed to be accurate to within 0.1 g cm^{-3}.

The variation of density with temperature can be estimated using the method of Grain (reported in [Nelken 1990]):

$$\rho_L = M\,\rho_{Lb}[3 - 2(T/T_b)]^n \quad (26)$$

where M = molecular weight, subscript "b" refers to the boiling point, and n = a constant that depends on chemical class (n = 0.25 for alcohols, 0.29 for hydrocarbons and 0.31 for other organics).

Abraham and McGowan [1987] reported a very simple method for the calculation of characteristic volume, which is closely correlated with molar volume. Atomic and bond contributions are: C 16.35, H 8.71, O 12.43, N 14.39, F 10.48, Cl 20.95, Br 26.21, I 34.53, S 22.91, P 24.87; for each bond, irrespective of type, subtract 6.56. Thus for NH_2COCH_3 the value is (2 x 16.35 + 12.43 + 14.39 + (5 x 8.71) – (8 x 6.56) = 50.59 cm^3 mol^{-1}; the experimental value of its molar volume is 50.86 cm^3 mol^{-1}.

There are five software programs that predict liquid density, namely ACDLabs, PREDICT, Molecular Modeling Pro, ProPred and SPARC. The ACDLabs website reports a standard error of 0.028 g cm^{-1} for the densities of a test set of 671 liquids. PREDICT is reported to yield errors of < 2%. The performance of the other software is not known.

It is recommended that one of the software programs or the Abraham and McGowan method [1987] be used for the calculation of liquid density and/or molar volume.

11. Prediction of viscosity of pure liquids

The viscosity data required under Annex VII of REACH are for aqueous solutions of chemicals. So far as can be ascertained, there are no methods available for the prediction of viscosity of aqueous solutions. This section therefore pertains only to the prediction of viscosity of pure liquids. Viscosity is important because it affects how a liquid should be handled, and also because it controls the permeation of liquids, for example in a spillage.

Liquid viscosity (η_L) can be regarded as a measure of the force needed to overcome the mutual attraction of molecules so that they can be displaced relative to each other [Grain 1990]. The prediction of liquid viscosity has been reviewed by Grain [1990] and Reinhard and Drefahl [1999].

The method of van Velzen et al [1972] is based on the following equation:

$$\log \eta_L = B_3(1/T - 1/T_0) \quad (27)$$

where B_3 = a class-dependent constant, and T_0 (in K) is the temperature at which the viscosity is 1 centipoise (cp). Grain [1990] gives details of how to calculate B_3.

Grain's method [Grain 1990] allows the calculation of viscosity at different temperatures, given the viscosity at the boiling point T_b (K):

$$\ln \eta_L = \ln \eta_{Lb} + B_4(1/T - 1/T_b) \quad (28)$$

Values of η_{Lb} are: alcohols and amines (aliphatic and aromatic) 0.45; all other organic liquids 0.2. The calculation of B_4 is given by Grain [1990].

Skubla [1985] developed a group contribution scheme for calculating the viscosity for various homologous series:

$$\log \eta_L = a_0 + a_1 P_{vap} \quad (29)$$

where a_0 and a_1 are derived from group contributions, and P_{vap} is vapour pressure.

Suzuki et al [1997] developed a 9-descriptor QSPR to model the viscosities of 237 organic liquids, with R^2 = 0.916 and RMS error = 0.167 log unit. When the same descriptors were used with a neural network model the statistics improved to R^2 = 0.958 and RMS error = 0.118 log unit. The descriptors, which were obtained from the ChemProp software [Schüürmann et al 1997], were not fully explained in the paper.

Katritzky et al [2000] used their CODESSA descriptors [www.semichem.com] to model the viscosities of a large diverse set of organic liquids:

$$\log \eta_L = 1.77\ \mathrm{HDCA}(2) + 0.000557\ G_I + N_{rings} + 20.2\ \mathrm{FPSA}(3) + 0.0897\ E_{min}(C) - 10.3 \qquad (30)$$

n = 361 R^2 = 0.854 s = 0.22

where HDCA(2) = hydrogen-bond donor charged surface area, G_I = gravitational index, N_{rings} = relative number of rings in a molecule, FPSA(3) = fractional positive partial charged surface area, and $E_{min}(C)$ = minimum atomic state energy for a carbon atom.

Kauffman and Jurs [2001] used their ADAPT software to develop a multiple linear regression QSPR for liquid viscosity, based on viscosity values (mPa.s) for a number of common organic solvents:

$$\log \eta_L = 0.263\ \mathrm{V0\text{-}1} + 0.0983\ \mathrm{DPOL\text{-}1} - 3.032\ \mathrm{SADH3} + 0.168\ \mathrm{SCDH{-}1} + 0.0710\ \mathrm{NRA\text{-}18} + 1.065\ \mathrm{FNSA\text{-}2} - 4.053\ \mathrm{FNSA\text{-}3} - 0.0681\ \mathrm{WPSA\text{-}3} - 1.475 \qquad (31)$$

n = 170 R^2 = 0.834 RMS error = 0.257 log unit

where V0-1= zero-order valence molecular connectivity, DPOL-1 = dipole moment, SADH-3 = (total surface area of all donatable H atoms)/(total molecular surface area), SCDH-1 = sum of (surface area x charge) for all donatable H atoms, NRA-18 =

number of ring atoms, FNSA-2 = (total charge-weighted partial negative surface area)/(total molecular surface area), FNSA-3 = (atomic charge-weighted partial negative surface area)/(total molecular surface area), and WPSA-3 = surface-weighted charged partial positive surface area.

Use of a neural network with the same descriptors gave much improved results (R^2 = 0.949, RMS error = 0.147 log unit). The ADAPT descriptors are available in the Pharma Algorithms ADME Boxes software.

Four software programs, namely ChemProp, Molecular Modeling Pro, PREDICT and ProPred, predict liquid viscosity. PREDICT is reported to yield errors of 2 – 20%, depending on the calculation method used. The performance of the other software is not known.

It is recommended that the method of van Velzen et al [1972] or of Grain [1990] be used for the estimation of liquid viscosity.

12. Prediction of surface tension of liquids

Surface tension of liquids affects, for example, leakage from a container and permeation into soils.

The prediction of surface tension (σ) has been reviewed by Grain [1990] and Reinhard and Drefahl [1999]. Grain's recommended method, which is applicable to diverse chemical classes, is the Macleod-Sugden approach [Macleod 1923, Sugden 1924], which uses the following equation:

$$\sigma = (P\rho_L/MW)^4 \qquad (32)$$

where P = parachor, ρ_L = liquid density, and MW = molecular weight. Parachor is an easily calculated additive property, and Grain [1990] lists incremental P values for numerous atoms, groups and bonds. Parachor can also be calculated by the ACDLabs

software. Reid et al [1977] found the Macleod-Sugden equation to yield a mean error in σ of 4.5%.

If the density of the liquid is unknown, the method of Grain [1990] can be used:

$$\sigma = [(\underline{P}/ V_b)(1 + k)(3 - 2T/T_b)^n]^4 \quad (33)$$

where k and n are constants listed by Grain [1990] for different chemical classes, V_b = molar volume at boiling point, T_b = boiling point (K). The method allows the calculation of surface tension at different temperatures, and has a mean error of 5.1%.

Another approach to the calculation of surface tension at different temperatures is given by the Othmer equation:

$$\sigma (T) = \sigma_{ref} [(T_c - T)(T_c - T_{ref})]^{11/9} \quad (34)$$

where the subscript "ref" refers to a reference temperature T_{ref}, and T_c = critical temperature. The parameters needed for the Othmer equation have been reported by Yaws et al [1991] for 633 chemicals.

Stanton and Jurs [1992], using their ADAPT software, developed a 10-descriptor model for a set of 146 alkanes, esters and alcohols, and found a standard error of 0.4 dyne/cm (mN/m). Kauffman and Jurs [2001] later developed a more general 8-descriptor model based on the surface tensions of 159 common organic solvents:

$$\sigma = - 3.301 \text{ KAPA-4} - 19.46 \text{ QPOS-1} + 87.40 \text{ SADH-3} + 0.0595 \text{ PPSA-1} - 140.6 \text{ FNSA-3} + 9.501 \text{ GRAV-3} + 0.200 \text{ RNCS-1} - 13.96 \text{ SAAA-3} - 46.80 \quad (35)$$

$n = 159 \quad R^2 = 0.835 \quad \text{RMS error} = 3.37 \text{ mN m}^{-1}$

where KAPA-4 = kappa shape index, QPOS-1 = charge on most positive atom, SADH-3 = (total surface area of all donatable H atoms)/(total molecular surface area), PPSA-1 = total surface area of all partial positively charged atoms, FNSA-3 = (atomic

charge-weighted partial negative surface area)/(total molecular surface area), GRAV-3 = cube root of gravitational index. RNCS-1 = relative negatively charged surface area, and SAAA-3 = (sum of surface area of acceptor atoms)/(total molecular surface area).

Use of a neural network with the same descriptors gave much improved results (R^2 = 0.931, RMS error = 2.22 mN m^{-1}). The ADAPT descriptors are available in the Pharma Algorithms ADME Boxes software.

Delgado and Diaz [2006] used a relatively large training set to derive a 6-term QSPR:

$$\sigma = 56.60\ N_C^R + 48.40\ N_O^R + 83.09\ N_N^R + 0.98\ MW + 3.47\ {}^3\chi^v + 0.16\ HDSA\text{-}1 - 6.45 \quad (36)$$

$n = 320 \quad R^2 = 0.96 \quad s = 1.43$ dyn cm^{-1}

where N_X^R = relative number of X atoms, MW = molecular weight, ${}^3\chi^v$ = 3rd order valence molecular connectivity, and HDSA-1 = hydrogen-bond donor surface area. The relative descriptors were obtained by divding the number of X atoms by the total number of atoms in the molecule. A test set of 55 compounds yielded r^2 = 0.94, s = 1.52 dyn cm^{-1}.

There are three software programs available that calculate surface tension. For a test set of 432 liquid compounds, the ACD/Labs software gave a standard error of prediction of 2.84 dyn cm^{-1}. PREDICT is reported to yield errors of about 5%. The performance of ProPred is not known.

It is recommended that the McLeod-Sugden method [Macleod 1923, Sugden 1924] or the Grain method [Grain 1990], or one or more of the above software programs, be used for the prediction of surface tension. The computer-based methods are especially useful if predictions are required for large numbers of chemicals.

13. Prediction of flash point

There have been very few publications on QSPR prediction of flash point. Hagopian [1990] and Vidal et al [2004] have reviewed the published work.

Butler et al [1956] reported a correlation of flash point determined by the closed cup method (T_F, ^oF) with boiling point (T_B, ^oF) for hydrocarbons:

$$T_F = 0.683\ T_B - 119 \qquad (37)$$

No standard error was given, although it may be noted that of 29 chemicals used to derive the above QSPR, 24 were predicted within ± 14^{o}F (8^{o}C) of the measured value; the maximum error was 40^{o}F (22^{o}C) for tetralin.

Hagopian [1990] used a similar approach to model the flash points of alcohols, aldehydes, amines and ketones, determined by both closed and open cup methods, with a different QSPR for each. For example, for alcohols he reported the following QSPRs and the mean absolute errors (MAE) of prediction:

Closed cup

$$T_F = 0.706\ T_B - 77.1 \qquad (38)$$

n = 31 MAE = 6.7^{o}F

Open cup

$$T_F = 0.688\ T_B - 66.4 \qquad (39)$$

n = 25 MAE = 7.4^{o}F

Satyanarayana and Rao [1992] also modelled the flash point of a large number of diverse organic chemicals using boiling point, with different correlations for different chemical classes. Using a test set of 1221 compounds they found the mean absolute error to be < 1%.

Hshieh [1997] used a quadratic equation in T_B (^{o}C) to model the flash points of a large number of diverse organic chemicals:

$$T_F = -\ 51.24 + 0.499\ T_B + 0.00047\ T_B^{\,2} \qquad (40)$$

n = 494 $R^2 = 0.935$ s = 11.7°C

Suzuki et al [1991] used principal components analysis to correlate the flash points of 400 diverse chemicals with two structural factors related to 1st order molecular connectivity ($^1\chi$) and the polar characteristics (G_i) of the functional groups:

$$T_F\ (^oC) = 25.57\ ^1\chi + \Sigma n_i G_i - 86 \qquad (41)$$

N = 400 $R^2 = 0.935$ s = 13.5°

Tetteh et al [1999] expanded this work by the use of a radial function neural network analysis, and were able to reduce the mean absolute prediction errors to 10-12°.

Katritzky et al [2001] used their CODESSA descriptors [www.semichem.com] to model flash points of a diverse set of 271 chemicals:

$$T_F = 44.5\ G_b^{1/3} + 16731\ HDCA + 4.95\ MW_R - 117.7 \qquad (42)$$

n = 271 $R^2 = 0.902$ s = 16.1°

where G_b = gravitational index over all bonded atoms, HDCA = charged solvent-accessible surface area of donatable H atoms, and MW_R = relative molecular weight.

By incorporating measured boiling point they were able to reduce the standard error to 11.2°, and by incorporating calculated boiling point the standard error was 14.2°.

Zhokhova et al [2003] critically analysed the work of Tetteh et al [1999] and Katritzky et al [2001], showing that there were some database errors in those publications. Zhokhova et al [2003] used a fragmental approach to model flash point. They developed several QSPR equations that gave good predictions. Thus, for a training set of 266 diverse compounds and using 9 indicator descriptors (counts of numbers of 9 different specified molecular fragments), they obtained $R^2 = 0.872$ and s = 18.8°. They obtained improved models with additional indicator descriptors, but did not report the QSPRs.

Gramatica et al [2004] used two Dragon descriptors [www.virtuallaboratory.org/lab/edragon] to model a small set of 35 esters with R^2 = 0.926 (s not given).

It appears from the above that a molecular fragment approach is the simplest and best way currently to model flash point. Of course such an approach means that predictions cannot be made for compounds that do not contain the molecular fragments used to train the model.

Only two software programs, namely ACDLabs and ProPred, appear to estimate flash point. However, no indications of performance are given.

It is recommended that either the approach of Hshieh [1997], Zhokhova et al [2003] or the ACDLabs software be used to predict flash point. Boiling points can be estimated as indicated in section 6, if experimental values are not available.

14. Prediction of auto-ignition temperature

There have been only a very few publications concerning the prediction of auto-ignition temperature (AIT). Taskinen and Yliruusi [2003] have reviewed the available literature. Tetteh et al [1996, 1998] used radial basis function neural networks to model a set of 232 organic chemicals with 13 different functional groups. They obtained a mean test set error of 33°.

Egolf and Jurs [1992] used their ADAPT software to model AIT values of hydrocarbons, alcohols and esters. They had to develop a separate QSPR for each class of chemical. For alcohols, for example, they found:

$$T_{AIT}\,(K) = -18.69\,NSB + 2640\,VCC7 + 152.2\,BJI - 12.80\,STS + 395.1 \qquad (43)$$

$n = 28 \quad R^2 = 0.941 \quad s = 24^o$

where NSB = number of single bonds, VCC7 = valence chain-7 molecular connectivity, BJI = Balaban J-index, and STS = steric strain.

Mitchell and Jurs [1997] used their ADAPT software to model AIT values of a data set of 327 diverse organic chemicals. They were unable to obtain good correlations for the whole data set, but found improved correlations when the chemicals were divided into hydrocarbons, nitrogen compounds, oxygen/sulphur compounds and alcohols/ethers. For example, for alcohol/ether compounds they obtained:

$$T_{AIT}\,(K) = -1700\,QPOS + 820\,RPCG + 1.94\,SAAA + 236\,RDTA - 197\,V4P + 43.4\,N2P + 136 \qquad (44)$$

$n = 67 \quad R^2 = 0.854 \quad \text{RMS error} = 35.0^{o}$

where QPOS = charge on most positive atom, RPCG = relative positive charge, SAAA = sum of surface areas of hydrogen bonding acceptor groups, RDTA = ratio of number of hydrogen bonding donor groups to hydrogen bonding acceptor groups, V4P = fourth order valence path molecular connectivity, and N2P = count of 2nd order paths. The authors commented that their prediction errors were in the range of experimental errors. The ADAPT descriptors are available in the Pharma Algorithms ADME Boxes software.

None of the approaches given above is very amenable to general use, so it is unfortunately the case that there is at present no simple method available for the prediction of auto-ignition temperature. The Mitchell and Jurs [1997] method is perhaps the least difficult of those mentioned above, provided that the ADAPT descriptors are accessible (ADMEWORKS Predictor from www.fqs.pl).

15. Prediction of soil sorption

Soil sorption involves the take-up of chemicals, usually into the organic surface coating of soil particles, from the surrounding milieu (usually an aqueous phase). It is in effect a partitioning process, and soil sorption (K_{oc}) of organic non-ionic chemicals can be estimated from their octanol-water partition coefficient (K_{ow}), as well as from other properties such as aqueous solubility. The subscript "oc" stands for "organic carbon". A number of reviews of K_{oc} prediction have been published recently [Lyman

1990, Reinhard & Drefahl 1999, Doucette 2000, Delle Site 2001, Doucette 2003, Dearden 2004, Kahn et al 2005]. That of Doucette [2000] contains a number of worked examples of the estimation of log K_{oc} values.

Sabljić et al [1995] correlated log K_{oc} values of several chemical classes with log K_{ow} values, and obtained reasonably good correlations. They found, however, that slopes and intercepts varied widely from class to class. For example, for hydrocarbons and halogenated hydrocarbons the correlation was:

$$\log K_{oc} = 0.81 \log K_{ow} + 0.10 \qquad (45)$$
$$n = 81 \quad r^2 = 0.887 \quad s = 0.451$$

That for anilines was:

$$\log K_{oc} = 0.62 \log K_{ow} + 0.85 \qquad (46)$$
$$n = 20 \quad r^2 = 0.808 \quad s = 0.341$$

It might be thought that such differences should mean that good correlations could not be obtained for diverse data sets. In fact, that has been shown not to be the case. Gerstl [1990] found, for a large diverse data set, a correlation as good as most of those of Sabljić et al [1995]:

$$\log K_{oc} = 0.679 \log K_{ow} + 0.094 \qquad (47)$$
$$n = 419 \quad r^2 = 0.831 \quad s \text{ not given}$$

Briggs [1981] found a good correlation for a large set of pesticides:

$$\log K_{om} = 0.53 \log K_{ow} + 0.64 \qquad (48)$$
$$n = 105 \quad r^2 = 0.90 \quad s \text{ not given}$$

Note that the soil sorption term here is K_{om}, where "om" stands for "organic matter". The relationship between K_{oc} and K_{om} is: log K_{oc} = log K_{om} + 0.2365 [Nendza 1998].

Hence from the two correlations above an estimate of K_{oc} or K_{om} can readily be obtained. Calculated log K_{ow} values can quickly be obtained from, for example, the ChemSilico website [www.logp.com].

The Abraham descriptors have been used [1999] to model K_{oc} values of a large diverse data set:

$$\log K_{oc} = 0.74\, R - 0.31\, \Sigma\alpha^{H} - 2.27\, \Sigma\beta^{O} + 2.09\, V_X + 0.21 \qquad (49)$$

$$n = 131 \qquad R^2 = 0.955 \qquad s = 0.245$$

where R = excess molar refractivity (a measure of polarisability), $\Sigma\alpha^{H}$ = hydrogen bond donor ability, $\Sigma\beta^{O}$ = hydrogen bond acceptor ability of oxygen, and V_X = McGowan molecular volume. The descriptors are approximately autoscaled, so that the magnitude of each coefficient is an indication of the relative contribution of each descriptor to soil sorption. Hence hydrogen bond acceptor ability and molecular size appear to be the most important factors controlling soil sorption. The Abraham descriptors can be calculated using the Absolv-2 software.

Kahn et al [2005] used log K_{ow} and CODESSA descriptors [www.semichem.com] to model soil sorption of a large diverse data set:

$$\log K_{oc} = 0.424 \log K_{ow} + 0.00272\ \text{PNSA-1} - 0.241\, \eta - 0.404\, P_{\pi-\pi}^{max} + 2.156 \qquad (50)$$

$$n = 344 \qquad R^2 = 0.759 \qquad s = 0.409$$

where PNSA-1 = partial negative surface area, η = absolute hardness, and $P_{\pi-\pi}^{max}$ = maximum π–π bond order.

Tao et al [1999] used a combination of 74 fragmental constants and 24 structural factors to model soil sorption of 592 diverse organic chemicals, with a standard error of 0.366 log unit. Although this is a good prediction, fragmental constant methods are not always easy to use, and can be tedious.

Delgado et al [2003] developed a very simple group contribution approach to model soil sorption, albeit with a relatively small training set:

$$\log K_{oc} = 0.60\, N_{\phi} + 0.0102\ \mathrm{MW} - 0.48\, N_N - 0.25\, N_O + 0.61\, N_S + 0.51 \qquad (51)$$

$n = 82 \quad R^2 = 0.94 \quad s = 0.33$

where N_{ϕ} = number of benzene rings, MW = molecular weight, N_N = number of nitrogen atoms, N_O = number of oxygen atoms, and N_S = number of sulphur atoms. A validation set of 43 chemicals yielded $R^2 = 0.96$, $s = 0.30$.

Although soil sorption varies to some extent with temperature, there do not appear to be any QSPR studies concerning this. One study has been published concerning the effect of ionisation on K_{oc} values. Bintein and Devillers [1994] reported the following QSPR based on 229 data points for 53 diverse chemicals:

$$\log K_p = 0.93 \log K_{ow} + 1.09\, f_{oc} + 0.32\ \mathrm{CFa} - 0.55\ \mathrm{CFb'} + 0.25 \qquad (52)$$

$n = 229 \quad R^2 = 0.933 \quad s = 0.433$

where K_p = sorption coefficient uncorrected for organic content, f_{oc} = fraction of organic carbon in soil, Cfa = correction factor for acid ionisation, and CFb' = correction factor for base ionisation.

Bearing in mind the large experimental error associated with soil sorption measurements [Nendza 1998], the standard errors given above are as good as can be hoped for.

There are three software programs that calculate log K_{oc} values. Using a test set of 100 diverse organic chemicals, Dearden [2004] compared the performance of two of them, and the results are shown below. The performance of the Pharma Algorithms ADME Boxes is not known.

Table 12. Software performance for prediction of soil sorption of a 100-compound test set

Software	% Predicted within ± 0.5 log unit of measured value	Mean absolute error (log unit)
Episuite	82%	0.490
Absolv-2	70%	0.569

It is recommended that the QSPRs developed by Gerstl [1990] or by Delgado et al [2003] and the Episuite software be used for estimation of log K_{oc}.

16. References

Abraham M.H., Andonian-Haftvan J., Whiting G.S., Leo A. and Taft R.S. (a) Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapours in water at 298 K, and a new method for its determination. *J. Chem. Soc. Perkin Trans. 2*, (1994) 1777-1791.

Abraham M.H., Chadha H.S. and Mitchell R.C. (b) Hydrogen bonding. 32. An analysis of water-octanol and water-cyclohexane partitioning and the log P parameter of Seiler. *J. Pharm. Sci.*(1994) **83**, 1085-1100.

Abraham M.H. and Le J. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.* (1999) **88**, 868-880.

Abraham M.H. and McGowan J.C. The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography. *Chromatographia* (1987) **23**, 243-246.

Abramowitz R. and Yalkowsky S.H. Melting point, boiling point and symmetry. *Pharm. Res.* (1990) **7**, 942-947.

Abramowitz R. and Yalkowsky S.H. Estimation of aqueous solubility and melting point of PCB congeners. *Chemosphere* (1990) **21**, 1221-1229.

Abshear T., Banik G.M., D'Souza M.L., Nedwed K. and Peng C. A model validation and consensus building environment. *SAR QSAR Environ. Res.* (2006) **17**, 311-321.

Andreev N.N., Kuznetsov S.E. and Storozhenko S.Y. Prediction of vapour pressure and boiling points of aliphatic compounds. *Mendeleev Commun.* (1994), 173-174.

Anliker R. and Moser P. The limits of bioaccumulation of organic pigments in fish: their relation to the partition coefficient and the solubility in water and octanol. *Ecotoxicol. Environ. Safety* (1987) **13**, 43-52.

Banks W.H. Considerations of a vapour pressure-temperature equation, and their relation to Burnop's boiling point function. *J. Chem. Soc.* (1939) 826-829.

Barratt M.D. QSAR, read-across and REACH. *ATLA* (2003) **31**, 463-465.

Basak S.C. and Mills D. Use of mathematical structural invariants in the development of QSPR mdels. *Commun. Math. Comput. Chem.* (2001) No. 44, 15-30.

Bergström C.A.S., Norinder U., Luthman K. and Artursson P. Molecular descriptors influencing melting point and their role in classification of solid drugs. *J. Chem. Inf. Comput. Sci.* (2003) **43**, 1177-1185.

Bintein S. and Devillers J. QSAR for organic chemical sorption in soils and sediments. *Chemosphere* (1994) **28**, 1171-1188.

Bodor N, Gabanyi N.Z. and Wong C.-K. A new method for the estimation of partition coefficient. *J. Am. Chem. Soc.* (1989) **111**, 3783-3786.

Borman S. New QSAR techniques eyed for environmental assessments. *Chem. & Eng. News* (1990) **20** (2), 20-23.

Briggs G.G. Theoretical and experimental relationships between soil adsorption, octanol-water partition coefficients, water solubilities, bioconcentration factors and the parachor. *J. Agric. Food Chem.* (1981) **29**, 1050-1059.

Butler R.M., Cooke G.M., Lukk G.G. and Jameson B.G. Prediction of flash points of middle distillates. *Ind. Eng. Chem.* (1956) **48**, 808-812.

Cabani S., Gianni P., Mollica V. and Lepori L. Group contributions to the thermodynamic properties of non-ionic organic solutes in dilute aqueous solution. *J. Solut. Chem.* (1981) **10**, 563-595.

Constantinou L. and Gani R. New group contribution method for estimating properties of pure compounds. *Amer. Inst. Chem. Eng. J.* (1994) **40**, 1697-1710.

Dearden JC. Molecular structure and drug transport. In Ramsden CA (Ed.), *Comprehensive Medicinal Chemistry*, vol. 4, Pergamon Press, Oxford, 1990, pp. 375-411.

Davies M.N., Toseland C.P., Moss D.S. and Flower D.R. Benchmarking pKa prediction. *BMC Biochemistry* (2006) **7** (18), published on the web 2 June 2006; www.biomedcentral.com/1471-2091/7/18.

Dearden J.C. The prediction of melting point. In Charton M. and Charton B. (Eds.), *Advances in Quantitative Structure-Property Relationships*, Vol. 2, JAI Press, 1999, pp. 127-175.

Dearden J.C. Quantitative structure-property relationships for prediction of boiling point, vapour pressure, and melting point. *Environ. Toxicol. Chem.* (2003) **22**, 1696-1709.

Dearden J.C. QSAR modelling of soil sorption. In Cronin M.T.D. and Livingstone D.J. (Eds.), *Predicting Chemical Toxicity and Fate*, CRC Press, Boca Raton, FL, 2004, pp. 357-371.

Dearden J.C. *In silico* prediction of aqueous solubility. *Expert Opinion on Drug Discovery* **1** (2006) 31-52.

Dearden J.C. Unpublished work (2007).

Dearden J.C. and Bresnen G.M. Thermodynamics of water-octanol and water-cyclohexane partitioning of some aromatic compounds. *Int. J. Mol. Sci.* (2005) **5**, 119-129.

Dearden, J.C., Cronin, M.T.D., Ahmed, S.A. and Sharra, J.A. QSPR prediction of Henry's law constant: improved correlation with new parameters. In Gundertofte, K. and Jørgensen, F.S. (Eds), *Molecular Modeling and Prediction of Bioactivity.* Kluwer Academic/Plenum Publishers, New York, 2000, pp. 273-274.

Dearden J.C. and Cronin M.T.D. Quantitative structure-activity relationships (QSAR) in drug design. In Smith H.J. (Ed.), *Introduction to the Principles of Drug Design and Action*, 4th edition. CRC Press (Taylor & Francis), Boca Raton, FL, 2006, pp. 185-209.

Dearden J.C. and Lappin D.C. (2007). Unpublished information.

Dearden J.C., Netzeva T.I. and Bibby R. (a) A comparison of commercially available software for the prediction of partition coefficient. In Ford M., Livingstone D., Dearden J. and van de Waterbeemd H. (Eds.), *Designing Drugs and Crop Protectants: Processes, Problems and Solutions,* Blackwell, Oxford, 2003, pp. 168-169.

Dearden J.C., Netzeva T.I. and Bibby R. (b) A comparison of commercially available software for the prediction of aqueous solubility. In Ford M., Livingstone D., Dearden J. and van de Waterbeemd H. (Eds.), *Designing Drugs and Crop Protectants: Processes, Problems and Solutions.* Blackwell, Oxford, 2003, pp. 169-171.

Dearden J.C. and Schüürmann G. Quantitative structure-property relationships for predicting Henry's law constant from molecular structure. *Environ. Toxicol. Chem.* (2003) **22**, 1755-1770.

Delgado E.J., Alderete J.B. and Jaña G.A. A simple QSPR model for predicting soil sorption coefficients of polar and nonpolar organic compounds from molecular formula. *J. Chem. Inf. Comput. Sci.* (2003) **43**, 1928-1932.

Delgado E.J. and Diaz G.A. A molecular structure based model for predicting surface tension of organic compounds. *SAR QSAR Environ. Res.* (2006) **17**, 483-496.

Delle Site A. The vapour pressure of environmentally significant organic chemicals: a review of methods and data at ambient temperature. *J. Phys. Chem. Ref. Data* (1996) **26**, 157-193.

Delle Site A. Factors affecting sorption of organic compounds in natural sorbent/water systems and sorption coefficients for selected pollutants. *J. Phys. Chem. Ref. Data* (2001) **30**, 187-439.

de Roode D., Hoekzema C., de Vries-Buitenweg S., van de Waart B. and van der Hoeven J. QSARs in ecotoxicological risk assessment. *Regul. Toxicol. Pharmacol.* (2006) **45**, 24-35.

Devillers J., Domine D. and Karcher W. Estimating *n*-octanol/water partition coefficients from the autocorrelation method. *SAR QSAR Environ. Res.* (1995) **3**, 301-306.

Dimitrov S., Dimitrova G., Pavlov T., Dimitrova N., Patlewicz G., Niemela J. and Mekenyan O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* (2005) **45**, 839-849.

Doucette W.J. Soil and sediment sorption coefficients. In Boethling R.S. and Mackay D. (Eds.), *Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences.* Lewis, Boca Raton, FL, 2000, pp. 141-188.

Doucette W.J. Quantitative structure-activity relationships for predicting soil/sediment sorption coefficients for organic chemicals. *Environ. Toxicol. Chem.* (2003) **22**, 1771-1788.

EC (2006a). Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official Journal of the European Union* **L396/1** of 30.12.2006. OPOCE, Luxembourg. Available from: http://publications.europa.eu.

EC (2006b). Directive 2006/121/EC of the European Parliament and of the Council of 18 December 2006 amending Council Directive 67/548/EEC on the approximation of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances in order to adapt it to Regulation (EC) No 1907/2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) and establishing a European Chemicals Agency. *Official Journal of the European Union* **L 396/850** of 30.12.2006. OPOCE, Luxembourg. Available from: http://publications.europa.eu.

ECB (2007). Final report of REACH Implementation Project (RIP) 3.3-2. Technical Guidance Document to Industry on the Information Requirements for REACH. European Commission-Joint Research Centre, Ispra, Italy. Available from: http://ecb.jrc.it/REACH/RIP_FINAL_REPORTS.

ECETOC Technical Report No. 74: *QSARs in the Assessment of the Environmental Fate and Effects of Chemicals*. ECETOC, Brussels, 1998.

ECETOC Technical Report No. 89: *(Q)SARs: Evaluation of the Commercially Available Software for Human Health and Environmental Endpoints with Respect to Chemical Management Applications.* ECETOC, Brussels, 2003.

Egolf L.M. and Jurs P.C. Estimation of autoignition temperatures of hydrocarbons, alcohols, and esters from molecular structure. *Ind. Eng. Chem. Res.* (1992) **31**, 1798-1807.

Eriksson L., Jaworska J., Worth A.P., Cronin M.T.D., McDowell R.M. and Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspectives* (2003) **111**, 1361-1375.

Fujita T., Iwasa J. and Hansch C. A new substituent constant, π, derived from partition coefficients. *J. Amer. Chem. Soc.* (1964) **86**, 5175-5180.

Gerstl Z. Estimation of organic chemical sorption by soils. *J. Contaminant Hydrology* (1990) **6**, 357-375.

Ghose A.K., Pritchett A. and Crippen G.M. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships: III. Modeling hydrophobic interactions. *J. Comput. Chem.* (1988) **9**, 80-90.

Girolami G.S. A simple "back of the envelope" method for estimating the densities and molecular volumes of liquids and solids. *J. Chem. Educ.* 1994) **11**, 962-964.

Gironés X., Amat L., Robert D. and Carbó-Dorca R. Use of electron-electron repulsion energy as a molecular descriptor in QSAR and QSPR studies. *J. Comput.-Aided Mol. Design* (2000) **14**, 477-485.

Goll E.S. and Jurs P.C. Prediction of vapour pressures of hydrocarbons and halohydrocarbons from molecular structure with a computational neural network model. *J. Chem. Inf. Comput. Sci.* (1999) **39**, 1081-1089.

Grain C.F. Vapor pressure. In Lyman W.J., Reehl W.F. and Rosenblatt D.H. (Eds.), *Handbook of Chemical Property Estimation Methods*. American Chemical Society, Washington DC, 1990, pp. 14.1-14.20.

Grain C.F. Surface tension. In Lyman W.J., Reehl W.F. and Rosenblatt D.H. (Eds.), *Handbook of Chemical Property Estimation Methods.* American Chemical Society, Washington DC, 1990, pp. 20.1-20.16.

Grain C.F. Liquid viscosity. In Lyman W.J., Reehl W.F. and Rosenblatt D.H. (Eds.), *Handbook of Chemical Property Estimation Methods*. American Chemical Society, Washington DC, 1990, pp. 22.1-22.20.

Gramatica P., Battaini F. and Papa E. QSAR prediction of physicochemical properties of esters. *Fresenius Environ. Bull.* (2004) **13**, 1258-1262.

Grigoras S. A structural approach to calculate physical properties of pure organic substances: the critical temperature, critical volume and related properties. *J. Comput. Chem.* (1990) **11**, 493-510.

Hagiopan J.H. Flash points of pure substances. In Lyman W.J., Reehl W.F. and Rosenblatt D.H. (Eds.), *Handbook of Chemical Property Estimation Methods*. American Chemical Society, Washington DC, 1990, pp. 18.1-18.14.

Hall L.H. and Story C.T. Boiling point and critical temperature of a heterogeneous data set. QSAR with atom type electrotopological state indices using artificial neural networks. *J. Chem. Inf. Comput. Sci.* (1996) **36**, 1004-1014.

Hansch C. and Leo A. *Exploring QSAR: Fundamental and Applications in Chemistry and Biology.* American Chemical Society, Washington DC, 1995.

Hansch C., Quinlan J.E. and Lawrence G.L. The linear free-energy relationship between partition coefficients and aqueous solubility of liquids. *J. Org. Chem.* (1968) **33**, 347-350.

Harris J.C. and Hayes M.J. Acid dissociation constant. In Lyman W.J., Reehl W.F. and Rosenblatt D.H. (Eds.), *Handbook of Chemical Property Estimation Methods.* American Chemical Society, Washington DC, 1990, pp. 6.1-6.28.

Hawker D.W. and Connell D.W. Octanol-water partition coefficients of polychlorinated biphenyl congeners. *Environ. Sci. Technol.* (1988) **22**, 382-387.

Hilal S.H. and Karickhoff S.W. A rigorous test for SPARC's chemical reactivity models: estimation of more than 4300 ionisation pKa's. *Quant. Struct.-Act. Relat.* (1995) **14**, 348-355.

Hine J. and Mookerjee P.K. The intrinsic hydrophilic character of organic compounds. Correlations in terms of structural contributions. *J. Org. Chem.* (1974) **40**, 292-298.

Horvath A.L. *Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds.* Elsevier, Amsterdam, 1992.

Hshieh F. Note: correlation of closed-cup flash points with normal boiling points for silicone and general organic compounds. *Fire Mat.* (1997) **21**, 277-282.

Huuskonen J., Salo M. and Taskinen J. Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *J. Pharm. Sci.* (1997) **86**, 450-454.

Ivanciuc O., Ivanciuc T., Cabrol-Bass D. and Balaban A.T. Evaluation in quantitative structure-property relationship models of structural descriptors derived from information-theory operators. *J. Chem. Inf. Comput. Sci.* (2000) **40**, 631-643.

Joback K.G. and Reid R.C. Estimation of pure-component properties from group contributions. *Chem. Eng. Commun.* (1987) **57**, 233-243.

Kahn I., Fara D., Karelson M., Maran U. and Andersson P.L. QSPR treatment of the soil sorption coefficients of organic pollutants. *J. Chem. Inf. Model.* (2005) **45**, 94-105.

Karthikeyan M., Glen R.C. and Bender A. General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Model.* (2005) **45**, 581-590.

Kaschula C.H., Egan T.J., Hunter R., Basilico N., Parapini S., Taramelli D., Pasini E. and Monti D. Structure-activity relationships in 4-aminoquinoline antiplasmodials. The role of the group at the 7-position. *J. Med. Chem.* (2002) **45**, 3531-3539.

Katritzky A.R., Chen K., Wang Y., Karelson M., Lucic B., Trinajstic N., Suzuki T. and Schüürmann G. Prediction of liquid viscosity for organic compounds by a quantitative structure-property relationship. *J. Phys. Org. Chem.* (2000) **13**, 80-86.

Katritzky A.R., Maran U., Karelson M. and Lobanov V.S. Prediction of melting points for the substituted benzenes. *J. Chem. Inf. Comput. Sci.* (1997) **37**, 913-919.

(a) Katritzky A.R., Mu L., Lobanov V.S. and Karelson M. Correlation of boiling points with molecular structure. 1. A training of 298 diverse organics and a test set of 9 simple organics. *J. Phys. Chem.* (1996) **100**, 10400-10407.

(b) Katritzky A.R., Mu L. and Karelson M. A QSPR study of the solubility of gases and vapors in water. *J. Chem. Inf. Comput. Sci.* (1996) **36**, 1162-1168.

Katritzky A.R., Petrukhin R., Jain R. and Karelson M. QSPR analysis of flash points. *J. Chem. Inf. Comput. Sci.* (2001) **41**, 1521-1530.

Katritzky A.R., Wang Y., Sild S., Tamm T. and Karelson M. QSPR studies on vapour pressure, aqueous solubility, and the prediction of air-water partition coefficients. *J. Chem. Inf. Comput. Sci.* (1998) **38**, 720-725.

Kauffman G.W. and Jurs P.C. Prediction of surface tension, viscosity, and thermal conductivity for common organic solvents using quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* (2001) **41**, 408-418.

Kier L.B. and Hall L.H. *Molecular Connectivity in Chemistry and Drug Research*. Academic Press, San Diego, CA, 1976.

Kier L.B. and Hall L.H. *Molecular Structure Description: the Electrotopological State.* Academic Press, San Diego, CA, 1999.

Klamt A., Eckert F., Diedenhofen M. and Beck M.E. First principles calculations of aqueous pK(a) values for organic and inorganic acids using COSMO-RS reveal an inconsistency in the slope of the pK(a) scale. *J. Phys. Chem. A* (2003) **107**, 9380-9386.

Klopman G. and Fercu D. Application of the multiple computer automated structure evaluation methodology to a quantitative structure-activity relationship study of acidity. *J. Comput. Chem.* (1994) **15**, 1041-1050.

Klopman G. and Wang S. A computer automated structure evaluation (CASE) approach to calculation of partition coefficient. *J. Comput. Chem.* (1991) **12**, 1025-1032.

Klopman G. and Zhu H. Estimation of the aqueous solubility of organic molecules by the group contribution approach. *J. Chem. Inf. Comput. Sci.* (2001) **41**, 439-445.

Klopman G. and Zhu H. Recent methodologies for the estimation of n-octanol/water partition coefficients and their use in the prediction of membrane transport properties of drugs. *Mini-Rev. Med. Chem.* (2005) **5**, 127-133.

Kühne R., Ebert R.-U. and Schüürmann G. Estimation of vapour pressures for hydrocarbons and halogenated hydrocarbons from chemical structure by a neural network. *Chemosphere* (1997) **34**, 671-686.

Labute P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* (2000) **18**, 464-477.

Leo A. Octanol/water partition coefficients. In Boethling R.S. and Mackay D. (Eds.), *Handbook of Property Estimation Methods for Chemicals.* Lewis, Boca Raton, FL, 2000, pp. 89-114.

Leo A., Jow P.Y.C., Silipo C. and Hansch C. Calculation of hydrophobic constant (log P) from π and *f* values. *J. Med. Chem.* (1975) **18**, 865-868.

Liang C.K. and Gallagher D.A. QSPR prediction of vapour pressure from solely theoretically-derived descriptors. *J. Chem. Inf. Comput. Sci.* (1998) **38**, 321-324.

Livingstone D.J. Theoretical property predictions. *Current Topics in Med. Chem.* (2003) **3**, 1171-1192.

Lyman W.J. Octanol/water partition coefficient. In Lyman W.J., Reehl W.F. and Rosenblatt D.H. (Eds.), *Handbook of Chemical Property Estimation Methods*. American Chemical Society, Washington DC, 1990, pp. 1.1-1.54.

Lyman W.J. Solubility in water. In Lyman W.J., Reehl W.F. and Rosenblatt D.H. (Eds.), *Handbook of Chemical Property Estimation Methods*. American Chemical Society, Washington DC, 1990, pp. 2.1-2.52.

Lyman W.J. Adsorption coefficient for soils and sediments. In Lyman W.J., Reehl W.F. and Rosenblatt D.H. (Eds.), *Handbook of Chemical Property Estimation Methods*, American Chemical Society, Washington DC, 1990, pp. 4.1-4.33.

Lyman W.J. Boiling point. In Boethling R.S. and Mackay D. (Eds.), *Handbook of Property Estimation Methods for Chemicals*, Lewis, Boca Raton, FL, 2000, pp. 29-51.

Mackay D. Solubility in water. In Boethling R.S. and Mackay D. (Eds.), *Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences.* Lewis, Boca Raton, FL, 2000, pp. 125-139.

Mackay D., Shiu W.Y. and Ma K.C. Henry's law constant. In Boethling R.S. and Mackay D. (Eds.), *Handbook of Property Estimation Methods for Chemicals.* Lewis, Boca Raton, FL, 2000, pp. 69-87.

Macleod D.B. On a relation between surface tension and density. *Trans. Faraday Soc.* (1923) **19**, 38-42.

Mannhold R. and van de Waterbeemd H. Substructure and whole molecule approaches for calculating log P. . *Comput.-Aided Mol. Des.* (2001) **15**, 337-354.

Marrero J. and Gani R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilibr.* (2001) **183-184**, 183-208.

Meylan W.M. and Howard P.H. Bond contribution method for estimating Henry's law constants. *Environ. Toxicol. Chem.* (1991) **10**, 1283-1293.

Meylan W.M., Howard P.H. and Boethling R.S. Improved method for estimating water solubility from octanol/water partition coefficient. *Environ. Toxicol. Chem.* (1996) **15**, 100-106.

Miller M.M., Wasik S.P., Huang G.-L., Shiu W.Y. and Mackay D. Relationships between octanol-water partition coefficient and aqueous solubility. *Environ. Sci. Technol.* (1985) **19**, 522-529.

Mills E.J. On melting point and boiling point as related to composition. *Phil. Mag.* (Series 5) (1884) **17**, 173-187.

Mitchell B.E. and Jurs P.C. Prediction of autoignition temperatures of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* (1997) **37**, 538-547.

Modarresi H., Modarress H. and Dearden J.C. QSPR model of Henry's law constant for a diverse set of organic chemicals based on genetic algorithm-radial basis function network approach. *Chemosphere* (2007) **66**, 2067-2076.

Myrdal P., Manka A.M. and Yalkowsky S.H. AQUAFAC 3: aqueous functional group activity coefficients; application to the estimation of aqueous solubility. *Chemosphere* (1995) **30**, 1619-1637.

Nagy P., Novak K. and Szasz G. Theoretical calculations on the basicity of amines. 1. The use of molecular electrostatic potential for pKa prediction. *J. Mol. Struct. – Theochem* (1989) **60**, 257-270.

Nelken L.H. Densities of vapors, liquids and solids. In Lyman W.J., Reehl, W.F. and Rosenblatt D.H. (Eds.), *Handbook of Chemical Property Estimation Methods.* American Chemical Society, Washington DC, 1990, pp. 19.1-19.24.

Nendza M. *Structure-Activity Relationships in Environmental Sciences*. Chapman & Hall, London, 1998.

Netzeva T.I., Worth A.P., Aldenberg T., Benigni R., Cronin M.T.D., Gramatica P., Jaworska J.S., Kahn S., Klopman G., Marchant C.A., Myatt G., Nikolova-Jeliazkova N., Patlewicz G.Y., Perkins R., Roberts D.W., Schultz T.W., Stanton D.T., van de

Sandt J.J.M., Tong W., Veith G. and Yang C. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *ATLA* (2005) **33**, 155-173.

Niimi A.J. Solubility of organic chemicals in octanol, triolein and cod liver oil and relationships between solubility and partition coefficients. *Water Res.* (1991) **25**, 1515-1521.

Nikolova-Jeliazkova N. and Jaworska J. An approach to determining applicability domains for QSAR group contribution models: an analysis of SRC KOWWIN. *ATLA* (2005) **33**, 461-470.

Nirmalakhandan N.N. and Speece R.E. QSAR model for predicting Henry's constant. *Environ. Sci. Technol.* (1988) **22**, 1349-1357.

Nys G.G. and Rekker R.F. Statistical analysis of a series of partition coefficients with special reference to the predictability of folding of drug molecules. Introduction of hydrophobic fragmental constants (*f* values). *Chim. Ther.* (1973) **8**, 521-535.

OECD. Application of structure-activity relationships to the estimation of properties important in exposure assessment. Envioronment Monograph No. 67, Paris, 1993.

Poole S.K. and Poole C.F. Chromatographic models for the sorption of neutral organic compounds by soil from air and water. *J. Chromatogr. A* (1999) **845**, 381-400.

Raevsky O.A., Raevskaja O.E. and Schaper K.-J. Analysis of water solubility data on the basis of HYBOT descriptors. Part 3. Solubility of solid neutral chemicals and drugs. *QSAR Comb. Sci.* (2004) **23**, 327-343.

Ran Y. and Yalkowsky S.H. Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* (2001) **41**, 354-357.

Rechsteiner C.E. Boiling point. In Lyman W.J., Reehl W.F. and Rosenblatt D.H. (Eds.), *Handbook of Chemical Property Estimations Methods*, American Chemical Society, Washington DC, 1990, pp. 12.1-12.55.

Reid R.C., Prausnitz J.M. and Sherwood T.K. *The Properties of Gases and Liquids.*McGraw-Hill, New York, 1977.

Reinhard M. and Drefahl A. *Estimating Physicochemical Properties of Organic Compounds.* Wiley, New York, 1999.

Rekker R.F. *The Hydrophobic Fragmental Constant*, Elsevier, Amsterdam, 1977.

Russell C.J., Dixon S.L. and Jurs P.C. Computer-assisted study of the relationship between molecular structure and Henry's law constant. *Anal. Chem.* (1992) **64**, 1350-1355.

Russom C.L., Breton R.L., Walker J.D. and Bradbury S.P. An overview of the use of quantitative structure-activity relationships for ranking and prioritizing large chemical

inventories for environmental risk assessments. *Environ. Toxicol. Chem.* (2003) **22**, 1810-1821.

Sabljić A., Güsten H., Verhaar H. and Hermens J. QSAR modelling of soil sorption. Improvements and systematics of log K_{oc} *vs.* log K_{ow} correlations. *Chemosphere* (1995) **31**, 4489-4514.

Sage M.L. and Sage G.W. Vapor pressure. In Boethling R.S. and Mackay D. (Eds.), *Handbook of Property Estimation Methods for Chemicals.* Lewis, Boca Raton, FL, 2000, pp. 53-65.

Sakuratani Y., Kasai K., Noguchi Y. and Yamada J. Comparison of predictivities of log *P* calculation models based on experimental data for 134 simple organic compounds. *QSAR Comb. Sci.* (2007) **26**, 109-116.

Sanghvi T., Jain N., Yang G. and Yalkowsky S.H. Estimation of aqueous solubility by the general solubility equation (GSE) the easy way. *QSAR Comb. Sci.* (2003) **22**, 258-262.

Satyanarayana K. and Rao P.G. Improved euation to estimate flash points of organic compounds. *J. Hazardous Materials* (1992) **32**, 81-85.

Schultz T.W., Hewitt M., Netzeva T.I. and Cronin M.T.D. Assessing applicability domains of toxicological QSARs: definition, confidence in predicted values, and the role of mechanisms of action. *QSAR Comb. Sci.* (2007) **26**, 238-254.

Schüürmann G., Kühne R., Kleint F., Ebert R.-U., Rothenbacher C. and Herth P. A software system for automatic property estimation from molecular structure. In Chen F. and Schüürmann G. (Eds.), *QSAR in Environmental Sciences – VII.* SETAC Press, Pensacola, FL, 1997, pp. 93-114.

Schüürmann G. and Rothenbacher C. Evaluation of estimation methods for the air-water partition coefficient. *Fresenius Environ. Bull.* (1992) **1**, 10-15.

Schwarzenbach R.P., Gschwend P.M. and Imboden D.M. *Environmental Organic Chemistry*. Wiley, New York, 1993.

Sedykh A.Y. and Klopman G. A structural analogue approach to the prediction of the octanol-water partition coefficient. *J. Chem. Inf. Model.* (2006) **46**, 1598-1603.

Sijm D.T.H.M., Schüürmann G., de Vries P.J. and Opperhuizen A. Aqueous solubility, octanol solubility, and octanol/water partition coefficient of nine hydrophobic dyes. *Environ. Toxicol. Chem.* (1999) **18**, 1109-1117.

Simamora P. and Yalkowsky S.H. group contribution methods for predicting the melting points and boiling points of aromatic compounds. *Ind. Eng. Chem. Res.* (1994) **33**, 1405-1409.

Skubla P. Prediction of viscosity of organic liquids. *Coll. Czech. Chem. Commun.* (1985) **50**, 1907-1916.

Soriano E., Cerdan S. and Ballesteros P. Computational determination of pK(a) values. A comparison of different theoretical approaches and a novel procedure. *J. Mol. Struct. – Theochem* (2004) **684**, 121-128.

Stanton D.T. and Jurs P.C. Computer-assisted study of the relationship between molecular structure and surface tension of organic compounds. *J. Chem. Inf. Comput. Sci.* (1992) **32**, 109-115.

Stein S.E. and Brown R.L. Estimation of normal boiling points from group contributions. *J. Chem. Inf. Comput. Sci.* (1994) **34**, 581-587.

Sugden S. The influence of the orientation of surface molecules on the surface tension of pure liquids. *J. Chem. Soc.* (1924) **125**, 1167-1189.

Suzuki T., Ebert R.-U. and Schüürmann G. Development of both linear and nonlinear methods to predict the liquid viscosity at 20^{o}C of organic compounds. *J. Chem. Inf. Comput. Sci.* (1997) **37**, 1122-1128.

Suzuki T., Ohtaguchi K. and Koide K. A method for estimating flash points of organic compounds from molecular structures. *J. Chem. Eng. Japan* (1991) **24**, 258-261.

Tao S., Piao H., Dawson R., Lu X. and Hu H. Estimation of the organic carbon normalized sorption coefficient (K_{oc}) using the fragment constant method. *Environ. Sci. Tech.* (1999) **33**, 2719-2725.

Taskinen J. and Yliruusi J. Prediction of physicochemical properties based on neural network modelling. *Adv. Drug. Delivery Rev.* (2003) **55**, 1163-1183.

Tesconi M. and Yalkowsky S.H. Melting point. In Boethling R.S. and Mackay D. (Eds.), *Handbook of Property Estimation Methods for Chemicals.* Lewis, Boca Raton, FL, 2000, pp. 3-27.

Tetko I.V., Bruneau P., Mewes H.-W., Rohrer D.C. and Poda G.I. Can we estimate the accuracy of ADMET predictions? *Drug Discovery Today* (2006) **11**, 700-707.

Tetko I.V., Tanchuk V.Yu., Kasheva T.N. and Villa A.E.P. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* (2001) **41**, 1488-1493.

Tetteh J., Howells S., Metcalfe E. and Suzuki T. Optimization of radial basis function neural networks using biharmonic spline interpolation. *Chemometr. Intell. Lab. Syst.* (1998) **41**, 17-29.

Tetteh J., Metcalfe E., Howells S.L. Optimization of radial basis and backpropagation neural networks for modeling auto-ignition temperature by quantitative structure-property relationships. ***Chemometrics and Intelligent Laboratory Systems*** (1996) **32**, 177-191.

Tetteh J., Suzuki T., Metcalfe E. and Howells S. Quantitative structure-property relationships for the estimation of boiling point and flash point using a radial basis function neural network. *J. Chem. Inf. Comput. Sci.* (1999) **39**, 491-507.

Todeschini R., Vighi M., Finizio A. and Gramatica P. 3-D modelling and prediction by WHIM descriptors. Part 8. Toxicity and physico-chemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors. *SAR QSAR Environ. Res.* (1997) **7**, 173-193.

Topliss J.G. and Costello R.J. Chance correlations in structure-activity studies using multiple linear regression. *J. Med. Chem.* (1972) **15**, 1066-1069.

Topliss J.G. and Edwards R.P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* (1979) **22**, 1238-1244.

Tsakanikas P.D. and Yalkowsky S.H. Estimation of melting point of flexible molecules: aliphatic hydrocarbons. *Toxicol. Environ. Chem.* (1988) **17**, 19-33.

Tsantili-Kakoulidou A., Panderi I., Csizmadia F. and Darvas F. Prediction of distribution coefficient from structure 2. Validation of PrologD, and expert system. *J. Pharm. Sci.* (1997) **86**, 1173-1179.

Tu C.-H. Group-contribution method for the estimation of vapour pressures. *Fluid Phase Equilibr.* (1994) **99**, 105-120.

Tu C.-H. and Wu Y.-S. Group-contribution estimation of normal freezing points of organic compounds. *J. Chin. Inst. Chem. Eng.* (1996) **27**, 323-328.

van Velzen D., Lopes Cardozo R. and Lagenkamp H. A liquid viscosity-temperature-chemical constitution relation for organic compounds. *Ind. Eng. Chem. Fundam.* (1972) **11**, 20-25.

Vidal M., Rogers W.J., Holste J.C. and Mannan M.S. A review of estimation methods for flash points and flammability limits. *Process Safety Progress* (2004) **23**, 47-55.

Votano J.R., Parham M., Hall L.H., Kier L.B. and Hall L.M. Prediction of aqueous solubility based on large datasets using several QSPR models utilizing topological structure representation. *Chem. & Biodivers.* (2004) **11**, 1829-1841.

Walker J.D., Dearden J.C., Schultz T.W., Jaworska J. and Comber M.H.I. QSARs for new practitioners. In Walker J.D. (Ed.), *Quantitative Structure-Activity Relationships for Pollution Prevention, Toxicity Screening, Risk Assessment, and Web Applications.* SETAC Press, Pensacola, FL, 2003, pp. 3-18.

Walker J.D., Dimitrov S. and Mekenyan O. Using HPV chemical data to develop QSARs for non-HPV chemicals: opportunities to promote more efficient use of chemical testing resources. *QSAR Comb. Sci.* (2003) **22**, 396-395.

Walker J.D., Jaworska J., Comber M.H.I., Schultz T.W. and Dearden J.C. Guidelines for developing and using quantitative structure-activity relationships. *Environ. Toxicol. Chem.* **22** (2003) 1653-1665.

Walker J.D., de Wolf W. QSARs on the world wide web: a need for quality assurance to prevent misuse. In Walker J.D. (Ed.), *QSARs for Pollution Prevention, Toxicity Screening, Risk Assessment, and Web Applications.* SETAC Press, Pensacola, FL, 2003, pp. 199-201.

Yaffe D. and Cohen Y. Neural network based temperature-dependent quantitative structure property relationships (QSPRs) for predicting vapour pressure of hydrocarbons. *J. Chem. Inf. Comput. Sci.* (2001) **41**, 463-477.

Yalkowsky S.H. and Banerjee S. *Aqueous Solubility: Methods of Estimation for Organic Compounds.* Marcel Dekker, New York, 1992.

Yalkowsky SH, Valvani SC, Roseman TJ. (1983) Solubility and partitioning. 6. Octanol solubility and octanol-water partition coefficients. *J. Pharm. Sci.* 72, 866-870.

Yang G.-Y., Yu J., Wang Z.-Y., Zeng X.-L. and Ju X.-H. QSPR study on the aqueous solubility ($-\lg S_w$) and *n*-octanol/water partition coefficients ($\lg K_{ow}$) of polychlorinated dibenzo-*p*-dioxins (PCDDs). *QSAR Comb. Sci.* (2007) **26**, in press (published online 13 October 2006).

Yaws C.L., Yang H.-C. and Pan X. 633 Organic chemicals: surface tension data. *Chem. Eng.* (1991) March, 140-150.

Zhao L. and Yalkowsky S.H. A combined group contribution and molecular geometry approach for predicting melting points of aliphatic compounds. *Ind. Eng. Chem. Res.* (1999) **38**, 3581-3584.

Zhokhova N.I., Baskin I.I., Palyulin V.A., Zefirov A.N. and Zefirov N.S. Fragmental descriptors in QSPR: flash point calculations. *Russian Chem. Bull. Intl. Edn.* (2003) **52**, 1885-1892.

European Commission

EUR 23051 EN – Joint Research Centre – Institute for Health and Consumer Protection

Title: *In Silico* Prediction of Physicochemical Properties

Author(s): Dearden J and Worth A

Luxembourg: Office for Official Publications of the European Communities

2007 – 66 pp. – 21 x 29.7 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593

Abstract

This report provides a critical review of computational models, and in particular (quantitative) structure-property relationship (QSPR) models, that are available for the prediction of physicochemical properties. The emphasis of the review is on the usefulness of the models for the regulatory assessment of chemicals, particularly for the purposes of the new European legislation for the Registration, Evaluation, Authorisation and Restriction of CHemicals (REACH), which entered into force in the European Union (EU) on 1 June 2007.

It is estimated that some 30,000 chemicals will need to be further assessed under REACH. Clearly, the cost of determining the toxicological and ecotoxicological effects, the distribution and fate of 30,000 chemicals would be enormous. However, the legislation makes it clear that testing need not be carried out if adequate data can be obtained through information exchange between manufacturers, from in vitro testing, and from in silico predictions.

The effects of a chemical on a living organism or on its distribution in the environment is controlled by the physicochemical properties of the chemical. Important physicochemical properties in this respect are, for example, partition coefficient, aqueous solubility, vapour pressure and dissociation constant. Whilst all of these properties can be measured, it is much quicker and cheaper, and in many cases just as accurate, to calculate them by using dedicated software packages or by using (QSPRs). These *in silico* approaches are critically reviewed in this report.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.